The ATLAS Computing Model & Distributed Computing Evolution

Roger W L Jones for the ATLAS Collaboration[a]

[a] *Department of Physics, Lancaster University, Lancaster LA1 4YB, UK*

Abstract. Despite only a brief availability of beam-related data, the typical usage patterns and operational requirements of the ATLAS computing model have been exercised, and the model as originally constructed remains remarkably unchanged. Resource requirements have been revised, and cosmic ray running has exercised much of the model in both duration and volume. The operational model has been adapted in several ways to increase performance and meet the as-delivered functionality of the available middleware. There are also changes reflecting the emerging roles of the different data formats. The model continues to evolve with a heightened focus on end-user performance; the key tools developed in the operational system are outlined, with an emphasis on those under recent development.

Keywords: Computing model, distributed computing, Grid computing.
PACS: 07.05.-t, 89.20.Ff, 29.50.+v, 7.05.Wr

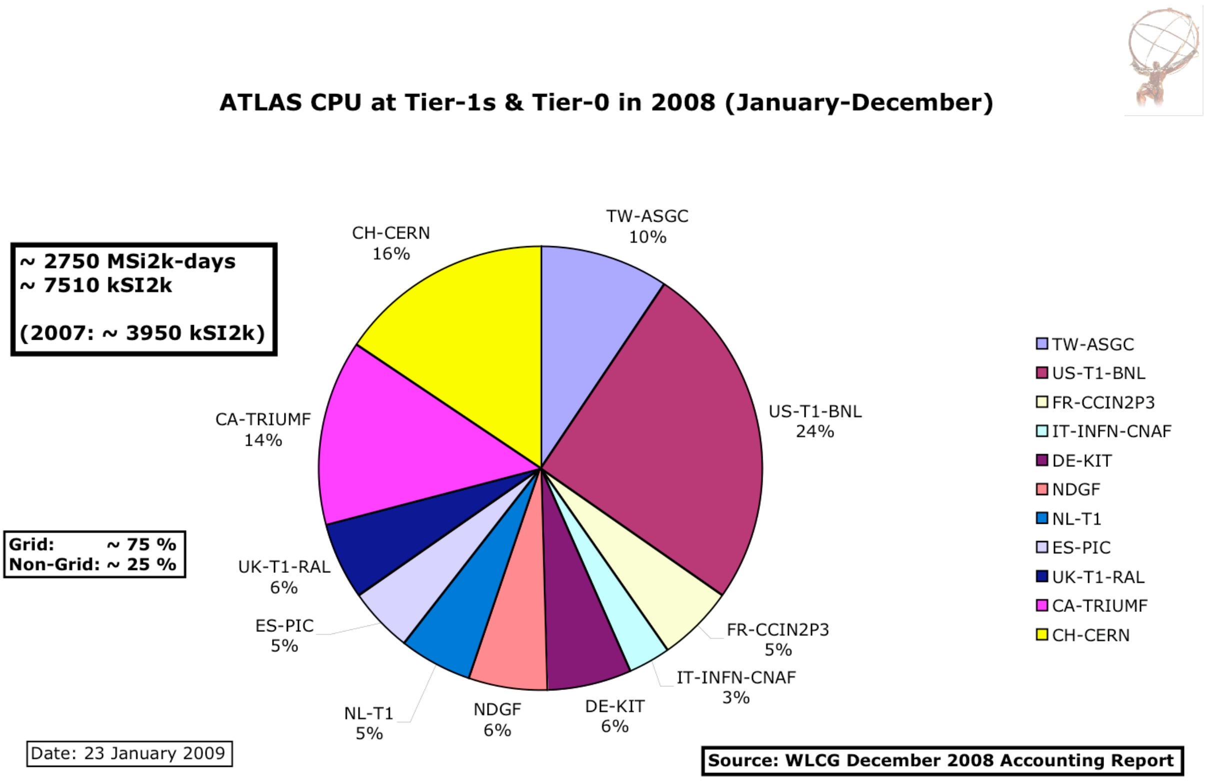
INTRODUCTION

The ATLAS computing model has been developed over more than ten years to meet the challenges of the LHC era. These challenges are many:

- Multi-petabyte sets of raw and processed data per year
- A rich set of representaions of each event, with both large and small files
- A global analysis community

These conditions lead, from the first, to a hierarchical model. As Grid technologies emerged, a refined version of the model was formed, with the Grid providing 'clouds' of resource that still had an underlying hierarchical structure. This model was presented to the Large Hadron Collider Committee in 2005 and a Computing Technical Design Report, C-TDR (1), produced. The model has experienced large scale tests through exercises with simulated data, since the autumn of 2006 with cosmic ray triggers and for a brief period in 2008 with beam-related data. All components have now received some level of testing, although full-scale analysis may still present surprises, as the scale and activity of the user base may change with real data. Through this process, the model has been adjusted to optimize the performance, overcome problems and to match the actual level of functionality delivered by the various Grid middleware.

The overall model has survived intact. The implementation has required more functionality to be developed within the experiment layer than had been expected initially. The resource requirements have been continually updated to reflect the changing running plans of the accelerator, the actual instantiation of the experiment event data model, the required level of simulation to match the understanding of the built detector response and the optimization of access to the datasets. The scale of the current deployment for is illustrated in Figure 1, which shows the break-down of processing used for ATLAS production tasks in 2008 at CERN and the Tier 1s, described below.



Source: WLCG December 2008 Accounting Report

.

FIGURE 1. The breakdown of processing used by ATLAS production in 2008 at CERN and the Tier 1s

THE TIER STRUCTURE AND ROLES

ATLAS still retains a three-tiered structure under central ATLAS control, with essentially the same roles as outlined in the C-TDR. Beyond these tiers, universities and groups have ‘Tier 3’ facilities. Various regional analysis centers are also emerging. The Tier structure is shown in Figure 2.

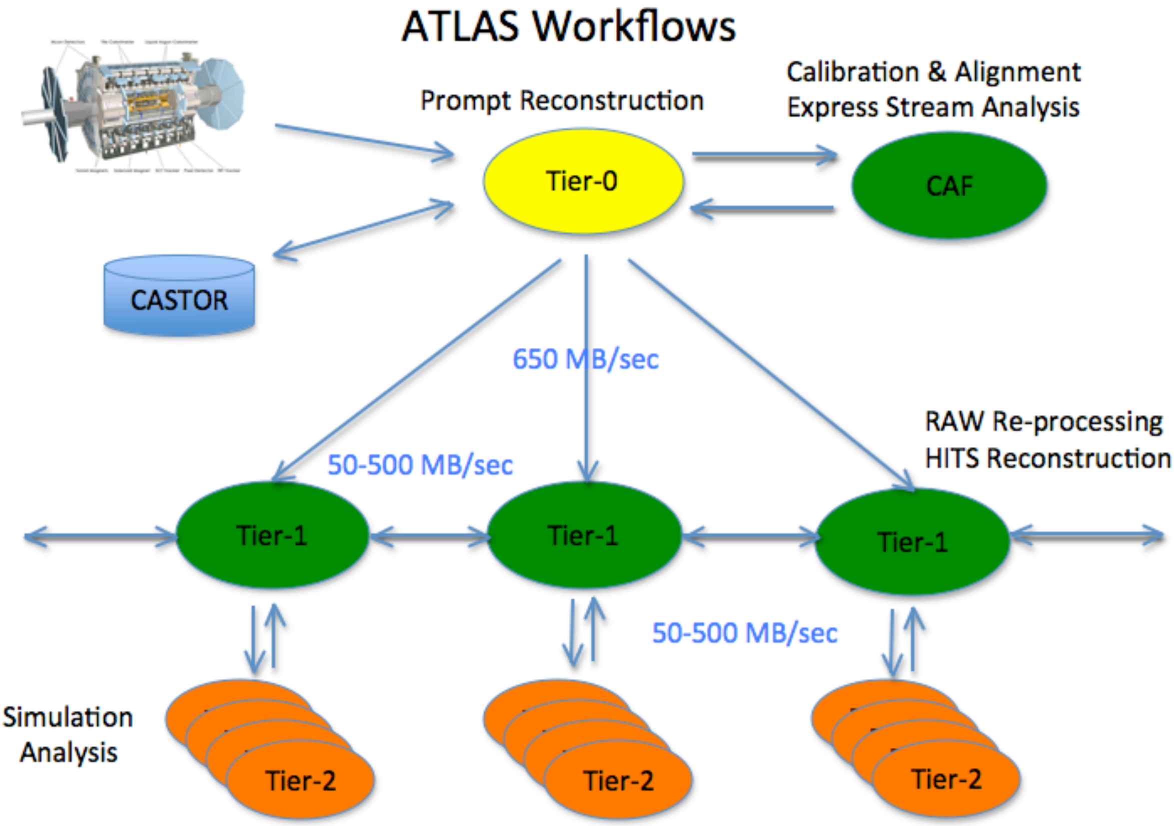


FIGURE 2. The centrally managed ATLAS Tier structure.

The CERN Facilities: Tier 0 and the Calibration and Alignment Facility

The model always had an important retained role for the Tier 0 at the accelerator laboratory, CERN. It performs the first pass processing of both express and calibration streams, and then of the bulk dataset. The initial estimates of the required disk capacity at CERN underestimated the required number of disk servers for the shipping of data in from the experiment and out to the other Tiers. As a consequence, the disk storage requirements are considerably greater than in early estimates, a lesson learned from the various scale exercises.

It was also evident from an early stage that additional facilities at CERN were needed for the initial calibration and alignment, as well as monitoring of the detector and to provide large-scale access to raw data for the preliminary understanding of the detector, trigger and reconstruction performance. CERN also hosts a CERN Calibration and Alignment Facility (CAF). The CAF capacity also includes the many servers needed for the basic operation of the experiment offline computing (such as build machines, machines for the automated software testing etc); this was not foreseen at the time of the C-TDR.

The Tier 1 Facilities

There are 10 Tier 1 facilities for ATLAS, generally hosted at national computing centers. These receive a portion of the raw data, and are responsible for its long-term curation and reprocessing to produce derived formats. In the steady state, this will occur twice a year: once after a few months, when better calibrations have been obtained; and once at the end of data taking, when algorithmic improvements can be applied. In the initial data-taking period, reprocessing be both faster and more frequent, with subsets of the data being processed of order ~5 times in a year. The Tier 1s also provide a facility for scheduled access by physics and detector performance groups to large quantities of data; this separation of scheduled activity from chaotic end-user analysis was planned from the start, to avoid interference with the higher-priority production tasks; this decision has been vindicated by experience. As the requirements for simulation have increased, the Tier 1s also provide part of the experiment simulation capacity.

The Tier 1s provide services for a cloud of associated Tier 2 facilities, which are usually based in universities. The Tier 1s host file catalogues and file transfer services, and act as an important staging point in the transfer of data around the world. The Tier 1s are linked to each other and to CERN by a dedicated Optical Private Network; the transfer of data from one Tier 1-Tier 2 cloud to another goes via the Tier 1s.

It was always seen as important that the Tier 1 production operations be separated from on-demand user access, as this can be very disruptive, especially to storage systems. However, there is a need for a well-defined subset of users to have limited read access to the data and some processing power for early data studies. As a consequence, 'power-user' access methods are being established at various sites. Their activities will be closely monitored to avoid disruptions to the processing activity.

The Tier 2 Facilities

The Tier 2 Facilities are where most of the users access the data, which is stored on Grid-visible disk. The usage is on-demand, and hence chaotic. An important 'down-scaling' from the initial model is that long-term user space is not provided at the Tier 2s. This is because the data management middleware does not yet support storage accounting and quotas on a per user basis. Instead, the model now provides a substantial scratch space for user output. Many files will be relatively short-lived, but those that must be kept for longer are either stored in a permanent storage areas that are managed by the physics and performance groups, or else the datasets are moved to private facilities. A third option exists in many clouds, where the country provides additional space on the Tier 2 or Tier 1 for uses from a geographical community. (It should be noted that where these analysis facilities are hosted at Tier 1 sites, a 'Chinese Wall' is required to ensure no disruption to the Tier 1 performance.) This 'regional' space is managed by members of that regional community.

The different storage areas are defined using the space token mechanism in the Storage Resource Manager (2), and access rights are controlled by role (e.g. production, user) and group (e.g. B-physics, UK-atlas). The intention is to restrict access to a small enough community as to make management relatively simple.

An important activity in the Tier 2 sites, especially during early data taking, will be the reconstruction of small samples of events with developing calibrations or algorithms. This requires access to the information conditions databases, which are hosted at the Tier 1s. Effective database access methods have been developed, including 'snapshots' that can be downloaded to the Tier 2 site, avoiding the need for remote database access.

DATA ORGANIZATION AND DISTRIBUTION

In the model described in the C-TDR, the analysis data formats were to be streamed according to their properties, allowing efficient access to the data; given the stream definition, the analyst need only read from suitable streams, not the whole dataset. The cost of this inclusive streaming is that this means some duplication of events between streams, and there is extra effort required to keep all processing versions consistent. In am important modification, the data is now not just streamed for analysis, but as it leaves the trigger system. The streams are defined by trigger types and are immutable under reprocessing. The argument for this is desire to reprocess some streams with high priority, given important changes in detector understanding or calibrations relevant to the stream. This also gives some improvement in the access patterns. The cost in terms of overlap of events between streams has been kept below the 10% level.

In another refinement of the original model, the data is also split into blocks where the specific luminosity of the beam (which controls the expected rate of interactions) is known and approximately constant. By requiring that subsequent actions are all acting on whole luminosity blocks, the integrated luminosity of the sample (essential for determining the probability of a given process to occur) may be calculated.

In a further attempt to optimize the management and access of the data, collections of files of similar data are grouped into datasets. The ATLAS data management system, DQ2 (3), then works with whole datasets, or in some cases collections of datasets. The latter mechanism allows similar data that is collected or processed at a later time to be grouped and accessed with earlier data without re-opening and changing the underlying dataset definitions.

The C-TDR model recognized that there would be a need for more data formats than the Raw-reconstructed-analysis chain planned in the central production. These additional formats were generically described as Derived Physics Data (DPD). This concept has proven to be very important. The derivations could be selections of subsets of data, the augmentation of the existing formats or the production of completely new and highly compact formats for specialized analyses. In the current planning, much of the DPD is actually to be made in the main processing step, and often consists of the selection of reconstructed format events for subsets of data important for detector calibration, algorithmic development or early physics studies. As such, many of the DPD sets in the current model correspond to the subsets of reconstructed data that were planned to be available on disk at the Tier 2s. Further derived formats will be made by physics and performance groups as part of their scheduled group analysis, and these will be stored in group areas in the Tier 2s. Full provenance information must be retained for these datasets. Another important development has been a subset of the analysis data in the ROOT (4) format that can be analyzed with tools from the main reconstruction suite.

Data distribution is essential for the operation of the model. The aim is to have most of the data required pre-placed, to avoid the strain on the system of many 'wildcat' data movements. The data in the Tier 2s is planned to have a typical useful lifetime of order several months. The use definition of several data streams aids with the distribution of the data, with sites within a Tier 2 cloud subscribing to pre-arranged streams. The production of derived formats also aids the process, with many following the associated stream into production spaces, while others are places in group-managed space at pre-arranged Tier 2s. The aim is to make all the data useful for user analysis available in each typical Tier 2 cloud.

User data movement and subscriptions are possible and required for small output sets. However, large data movements must be by pre-arrangement using official tools; policies are in place to enforce this, as large unplanned movements pose a threat to the stability of the system as a whole.

SIMULATION

Despite the absence of collision data from the Large Hadron Collider, the understanding of the 'as built' ATLAS detector has advanced considerably, mainly through the use of cosmic ray events. The simulation of the detector must reflect this understanding so that further advance can be made. Since the C-TDR, a full evaluation of the required physics performance and detailed geometry in the simulation means that the required processing capacity has been revised upwards. To help with this problem, a simulation mode - Atlfast II (5) - intermediate between the existing fast and full simulation modes has become increasingly important for many tasks. This produces analysis object format directly using an order of magnitude less processing power. This will be particularly important for studying background samples, but increases the importance of the validation of the full simulation with real data, as it is to this that the intermediate mode simulation is tuned. An even faster, more approximate, simulation mode is used for scoping and tuning studies.

THE OPERATIONAL DISTRIBUTED COMPUTING MODEL

Data management and Data Distribution

Despite the huge amount of data to be moved, bandwidth is not in general the major issue with data movement, although last-mile effects remain at some sites. The LHC Optical Private Network provides the required bandwidth between CERN and the Tier 1 sites, and national and transnational networks meet the Tier 1-Tier 2 connectivity requirements. However, the challenges to be met include the registration and cataloguing of the data, the rapid deletion of unwanted data, fault-tolerance and retries in the case of failures or partially successful transfers, location services for data already stored, and subscription mechanisms to automate the transfer of classes of data. To achieve all this, ATLAS is now using its second generation of Distributed Data Management(). This requires a careful balance between global services (such as repositories and location catalogues) and local site services such as local file catalogues. The global services run at CERN, the local file catalogues at the Tier 1s. The problem is simplified in several ways. The standard logical file names form the middleware are aggregated into datasets of related data and the associated metadata. These datasets are in turn aggregated into larger data collections. The global services deal with the datasets and larger aggregations, while the local services resolve the logical file names into physical file names on specific storage systems. The components are shown at high level and schematically in Figure 3.

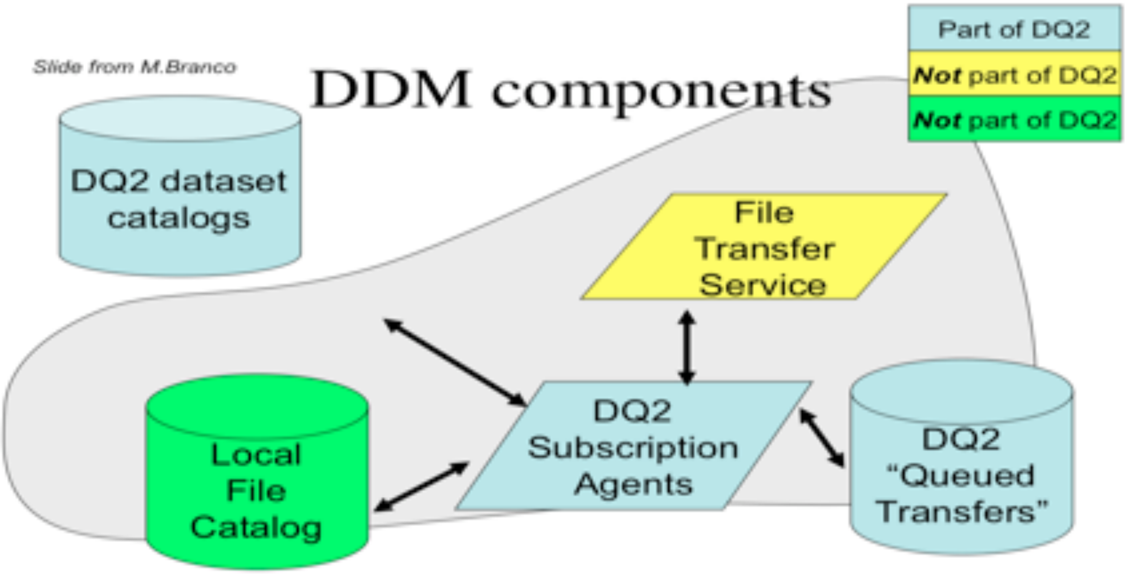


FIGURE 3. The Distributed Data Management System architecture

The ATLAS Production System

The production of simulated data and the processing and reprocessing of real and simulated data is handled by the production system. This takes datasets as inputs, transforms them through defined tasks and produces output datasets to be stored and registered. The job definition and processing information are stored in the production database, implemented in Oracle, and the running is supervised by the production system. The task request interface uses the Panda(6) system, with the monitoring and status information being aggregated and displayed by both the Panda and ATLAS dashboard systems; convergence between the two is underway. The whole system is based on the wLCG(7) middleware set, assembled and linked using software written by ATLAS.

Distributed Analysis

Distributed analysis is the most demanding activity, and two interfaces are in common use. GANGA is the most general, with back-ends to the EGEE(8), NorduGrid(9) and OSG(10) grid deployments, as well as local batch systems and local interactive running. It is designed to handle very general tasks, and is used by other communities. The pATHENA interface in Panda is devised to run tasks using the ATLAS Athena framework, and primarily interfaces to the Panda system. Again, convergence is planned between the two tools.

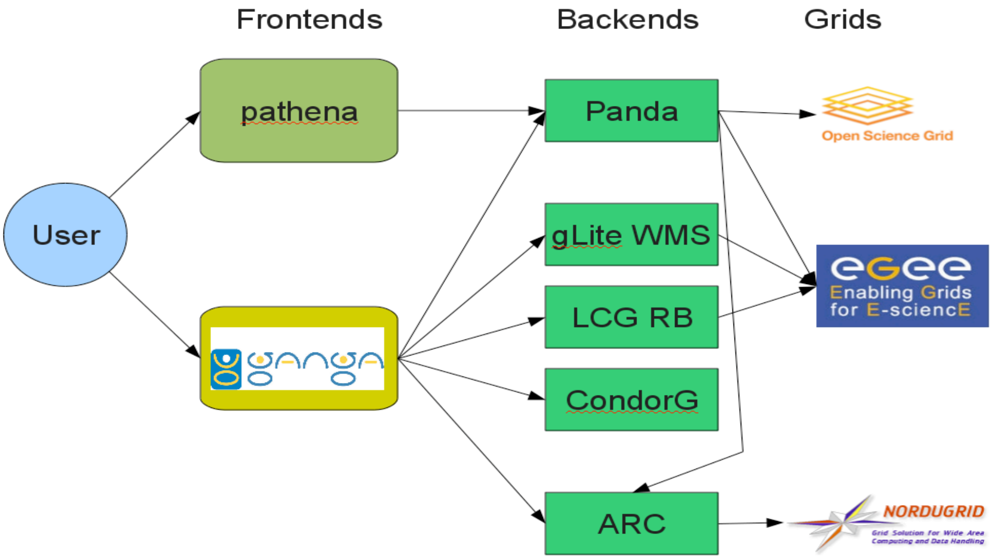

Frontends
Backends
Grids
User
pathena
ganga
Panda
gLite WMS
LCG RB
CondorG
ARC
Open Science Grid
egee Enabling Grids for E-sciencE
NORDUGRID Grid Solution for Wide Area Computing and Data Handling

FIGURE 4. Distributed Analysis in ATLAS

The ATLAS Grid Information System

While the middleware sets all provide some level of Grid information service, this is not sufficient for the needs of the experiment. Various information systems have evolved, but these are in the process of being replaced by a single ATLAS Grid Information System, AGIS (11). This aggregates the information from these older system, and receives new inputs from the task request interface, the Panda services, the data replication monitoring system, the logging system and the traditional grid information systems into one database with a web interface. This system will also include user information, such as the allowed privileges, roles and accounting information.

CONCLUSION

The main architecture of the ATLAS computing model is largely unchanged since its design in around 2005, although there have been changes to the required resources and some variation in the function at various tiers. The separation of activities into central, schedules production, semi-scheduled group analysis and chaotic user analysis remains, the chaotic element is divorced from the production system The model has been implemented by means of a distributed data management system to manage the files, aggregated into datasets. These are produced using the ATLAS production system and the distributed analysis tools. The system uses the ATLAS Grid Information System to share required information for each of the tasks and for the outcomes. There are several monitoring tools in place to keep track of the system, and convergence is underway between them.

The system has now been tested in large-scale challenges using simulated data, cosmic ray data read-out from the detector, and even real beam-related data from the detector. All components have been tested and seen to work at large scale. The immediate remaining challenges are sustained operation over a period of many months, and the growing demands of an expanding user-base. Into the future, the scaling to cope with growing data volumes will pose many challenges, but the existing system provides a good base form which to meet them.

ACKNOWLEDGMENTS

I would like to thank all in the ATLAS distributed computing, but in particular I should thank A Anisenkov, D Barberis, K Bos, M Branco, S Campana, A Farbin, J Elmsheuser, A Klimentov, D Krivashin, A Read, R Rocha, J Shank, A Vaniachine and T Wenaus for input into my talk and write-up.

REFERENCES

1. *The ATLAS Computing Technical Design Report*, ATLAS-TDR-017; CERN-LHCC-2005-022, June 2005.
2. *The Storage Resource Manager working group*, http://sdm.lbl.gov/srm-wg/index.html
3. *Managing ATLAS data on a petabyte-scale with DQ2*, M Branco et al 2008 J. Phys.: Conf. Ser. 119 062017
4. *ROOT* http://root.cern.ch/drupal/
5. *Atlfast II twiki* https://twiki.cern.ch/twiki/bin/view/Atlas/AtlfastII
6. *PanDA: Distributed production and distributed analysis system for ATLAS,* T. Maeno et al. s.l. : J.Phys.Conf.Ser.119:062036, 2008.
7. *The LHC computing grid project*. M. Lamanna et al. s.l. : NIM A, 2004, Vol. 534.
8. *Enabling Grids for E-sciencE*. [Online] http://www.eu-egee.org.
9. *The NorduGrid Project*. M. Ellert et al. s.l. : NIM A, 2003, Vol. 502.
10. *Open Science Grid*. [Online] http://www.opensciencegrid.org.
11. *Ganga: A tool for computational-task management and easy access to Grid resources*. J.T. Mościcki et al. 11, s.l. : Computer Physics Communications, 2009, Vol. 180.pATHENA
12. *AGIS*. [Online] http://dashb-build.cern.ch/build/unstable/doc/guides/agis/html/user/index.html