The ATLAS Computing Model

RWL Jones
Department of Physics, Lancaster University, Lancaster LA1 4YB, UK
E-mail: Roger.Jones@cern.ch

D Barberis
University of Genoa/INFN, Via Dodecaneso 33, IT-16146, Genoa, Italy
E-mail: Dario.Barberis@cern.ch

Abstract. The ATLAS Computing Model was constructed after early tests and was captured in the ATLAS Computing TDR in June 2005. Since then, the grid tools and services have evolved and their performance is starting to be understood through large-scale exercises. As real data taking becomes immanent, the computing model continues to evolve, with robustness and reliability being the watchwords for the early deployment. Particular areas of active development are the data placement and data access, and the interaction between the TAGs, the datasets and the Distributed Data Management issues. The earlier high-level policies and models are now being refined into lower level instantiations.

1. Introduction
ATLAS computing is now in the final phase before data-taking, and as such it is now concentrating on the pragmatic development of working systems that can sustain the expected early data rates and deliver physics output in a short time. This means that while many clever developments are still possible, and perhaps ultimately even needed, in the medium term the emphasis is on integration, deployment, testing, documentation and the development of an effective event data model. Correspondingly, ATLAS is also determining the simplest and most effective ideas that will work by confronting the model (described in full in the ATLAS Computing Technical Design Report [1]) with real-world data exercises and refining the details of the high-level plans. One of the most significant recent developments is that ATLAS is now taking real data using cosmic ray triggers, and the full chain of computing operations from the data acquisition to distributed analysis has been demonstrated in a single exercise. (An example cosmic ray event displayed with the ATLANTIS [2] event display is shown in figure 1). Considerable challenges remain, not least the reprocessing of data in the cloud of Tier 1 national centres. This and other components will be added in the Full Dress Rehearsals in Spring 2008.

1 To whom any correspondence should be addressed.
The various tiers of the ATLAS computing system have remained unchanged for some time. The Tier 0 is at CERN and performs monitoring, calibration and first-pass processing; it also hosts the first secure copy of the raw data and of the first pass output data formats. The 10 Tier 1 sites receive the raw data and the first pass processing, and host the second secure copy of the raw data. They provide scheduled access for physics and detector groups to the processed data, and perform the reprocessing when calibrations and algorithms improve. They also host copies of the conditions database. The Tier 2 sites are associated with a Tier 1, and provide on-demand access to processed data for physics analysts. They also provide the simulation capacity for the experiment. Tier 3 sites are less well defined, and are not available to the whole collaboration, although they may wish to make capacity available for simulation on a temporary basis. The Tier 3s can have many forms; they can be reserved capacity on Tier 1 or Tier 2 fabric, dedicated clusters for local analysis or else the traditional collection of desktop and laptop machines. A task force is currently formulating recommendations to Tier 3 sites to allow the best use of available resources. A common requirement for Tier 3 sites is that there be at least one Grid user interface to allow users to submit jobs to Tier 2s, and the ATLAS data movements need to be improved. The final Tier to be mentioned is the CERN Analysis Facility (CAF); this has some aspects of a Tier 2, but provides more access to the full reprocessed data (ESD) and the RAW data than would be normal, and is intended to allow calibration and algorithm development. As such, it will have a vital role in the early data taking.

The current ATLAS resource requirements are briefly summarized in table 1. We note that the pledges do not match the requirements in 2010, and even in 2008 there is a significant shortfall in projected disk capacity in the Tier 2s, which may have a very negative impact on physics analysis. The resource requirements beyond 2012 are now under active consideration.
Table 1: ATLAS projected resource requirements in 2008 and 2010.

|          | CPU (MSI2k) | Disk (PB) | Tape (PB) |
|----------|-------------|-----------|-----------|
|          | 2008 | 2010 | 2008 | 2010 | 2008 | 2010 |
| Tier 0   | 3.7  | 6.1  | 0.15 | 0.5  | 2.4  | 11.4 |
| CAF      | 2.1  | 4.6  | 1.0  | 2.8  | 0.4  | 1.0  |
| All Tier 1s | 18.1 | 50.0 | 10.0 | 40   | 7.7  | 28.7 |
| All Tier 2s | 17.5 | 51.5 | 7.7  | 22.1 |      |      |
| Total    | 41.4 | 112.2| 18.9 | 65.4 | 10.5 | 41.1 |

2. Roles and Permissions

One requirement for an effective Grid system that was identified early on by the experiments is the ability to grant permissions, monitor, account for resource usage and apply quotas with a finer resolution than simply the experiment/virtual organization. The simplest and most obvious distinction is between the production activity and that of the individual users. In the absence of this division, we have already observed that a user taking all of the resources, especially storage, can block production at sites. Brute force solutions have entailed a hard division between the production and user storage, which looses the flexibility the Grid initially offered. However, this is not enough. In an organization of nearly 2000 users, the resources must be managed in smaller subunits; there is the need for management in and by the group, with some members having distinct roles such as production or user. This ontology is supported by the Virtual Organization Membership Service (VOMS [3]), but as yet little of the middleware makes use of identities it supports. In particular, it is still not possible to allow storage quotas on this basis.

The above discussion has lead to a refinement of the ATLAS planning for disk space on Tier 2 sites. Groups will be given disk allocations in the Tier 2 cloud. This will be to accommodate their group derived physics data and also other derived data produced by group members and deemed to be important to the group or a subgroup. Such user datasets will originally reside in the substantial scratch area for ATLAS users on the Tier 2s; the intended lifetime on the scratch will be sufficient to allow files to be adopted, deleted or to be moved to Tier 3 sites. This scratch area will pose some technical challenges, as the clean-up of the space will have to be done locally but using ATLAS tools and following agreed policies. The requirement for ATLAS tools is that the user files will be registered in the ATLAS data management system, and so will have to be removed from the catalogues on deletion.

An associated refinement is that it is clear that the traffic between Tier 2 sites and Tier 3s will have to be managed, as high transfer rates will disrupt the Tier 2 operations for other users. Further, private transfers between any Tiers have the potential to disrupt the data management system. The rough policy will be that private user transfers at the level of tens of gigabytes a day pose no problem, but tens of terabytes a day will have to be throttled. This requires users to use ATLAS tools, and abuse to be met with sanctions.

The need for large-scale data movement by individual users should be largely reduced by the ATLAS policy of pre-placing and sending jobs to the data, not data to the jobs. This policy is starting to meet with some success, but was initially hampered both by problems with the data placement because of configuration and policy issues in the distributed data management. Another issue was user psychology; it has taken some time for users to understand that it is better to run a job at a remote site with data than attempt to hold everything locally. Fortunately, real data volumes are likely to quickly shatter such misconceptions.

All of the above discussion emphasizes the need for the proper integration of VOMS roles into all levels of middleware, and the ability to apply access control lists and quotas. It also implies accounting
and monitoring with the same level of refinement. We note with some satisfaction the improving
status of monitoring and accounting in the various Grid systems, but the VOMS roles are still to be
implemented.

3. Data Access Optimization
ATLAS is implementing two methods of data access optimization:

- Data is split into streams based in the trigger information about each event.
- Data collections are built using pointers contained in TAG datasets; each event contains an
  entry in the TAG that also carries about 100 words of other information describing the event.

Streams are largely to reduce the file opening, but can also allow selective and prioritized reprocessing
of subsets of the data. This may especially important in the early phase of data taking. By keeping
substreams within the original stream boundaries (for instance, when making analysis object data
(AOD) from the full reconstructed data (ESD)), back navigation is eased.

TAGs serve two purposes. The first is to allow direct access to events within files using pointer
information. The second is to define data collections using the stored event keyword information.
There are also two formats for the TAG.

- The first is a flat file (in Root [4] format), most suited to the event access role. Every
  reconstructed or analysis file in the system will have a corresponding fat file TAG, which may
  also be used for data collection building over small samples of data. This format has the
  advantage that quite complex arithmetic selections and queries may be made. Ongoing
  exercises envisage five primary physics streams, with other streams (express for monitoring,
  calibration, pathological events from the event filter etc) for technical developments. The
  initial streaming will be done at the event filter output, before the Tier 0.

- The second is a relational database format. This is suited for queries over large data sets. It has
  more limited query abilities, and will rapidly grow (by about 6TB a year). Such a database
  will require significant effort to maintain, and so will be limited to relatively few sites. Current
  tests are being made to establish how many database TAG instances are

4. Event data model
There is considerable effort going into the ATLAS event data model in various aspects. From the point
of view of performance and resources, there is a continuing effort to reduce event sizes and processing
times to the values required in the computing model. This is a moving picture; currently, the major
problem is with the simulation processing time.

Another major area concerns derived physics dataset (DPD) formats. Recent developments have
allowed some level of interoperability between the ATLAS framework (ATHENA) and Root, but
other formats are also in play. The best solution may depend on the number of events to be handled
and the information required in a particular analysis. While there are obvious advantages to a common
format, it is likely that some diversity will remain.

One area of DPD development that is tailoring the event data model to particular analyses involves
reducing the information in the AOD to produce a more compact format. Three options are available:

- Skimming: this implies the selection of the desired events from a larger dataset
- Thinning: this implies the selection of only the required objects or containers of objects from
  the desired events
- Slimming: this implies the selection of just the desired properties of an object.

The aim is to provide sufficient compression and flexibility that a small set of event formats can
satisfy the physics community, thus easing the burden of maintenance and encouraging cross-group
analysis.
5. Commissioning and tests

5.1. Cosmic ray running and Tier 0/Tier 1 transfer tests

ATLAS has been able to take some cosmic ray data since September 2006, and in 2007 and 2008 is running a series of data-taking periods for the full detector using cosmic ray triggers. The data is read out, copied to the Tier 0, processed and distributed to the Tier 1 centres, as it will be for the collision data. The full rate of Tier 0 processing has been achieved (both in these and in stand-alone tests using simulated data). The data has been exported and stored at half of the Tier 1s in cosmic tests and all in stand-alone exercises. In general, data has not yet been replicated onwards to the Tier 2s. Despite this, quasi-real time analysis has been performed in at least one Tier 2. A major milestone was the running of the full chain from DAQ to data analysis in the last week of August 2007.

Figure 2 shows the throughput and error rates achieved by the Tier 0 processing system in a Tier 0/Tier 1 processing transfer test. We are meeting and exceeding the required throughput.

Figure 2: The throughput and performance achieved in recent transfer tests between the Tier 0 and the Tier 1 centres.

5.2. Full dress rehearsals (FDRs)

Starting in 2001, ATLAS has been running large-scale exercises with simulated data, and since 2003 this has been using Grid technologies. The latest and last phase of these tests will be the Full Dress Rehearsals in which all the major components required for real data operations will be run with a realistic mix of data and triggers. This will include streaming of the data (the cosmic ray tests do not have the variety of triggers to make this practical there) and importantly the reprocessing of data at the Tier 1s. Another vital part will be the exercise of the rapid calibration procedures required to allow for the first-pass processing of the full dataset collected within a couple of days of arrival at the Tier 0. (Subsets of the data will be processed immediately for monitoring and calibration using older calibrations.)

An important goal of the FDRs will be to exercise the full range of analysis. To do this, some data will be pre-loaded into the Tier 1 and Tier 2 sites. Root-based analysis, trigger-aware analysis using the conditions and trigger databases and analyses of data from which the truth information has been stripped will all be performed in parallel with simulation and reconstruction tasks.
5.3. Common computing readiness challenges (CCRC)
Another important task will be to run the experiment computing models in parallel with those of the other experiments. ATLAS has already attempted to identify possible issues and conflicts, but real world exercises will also test the capacities of the computing centres, networks and firewalls under the full LHC load. Two CCRC exercises are planned, the first in February and the second in May 2008. These valuable additions to the programme unfortunately cause some conflict with the planned FDR exercises in ATLAS. As a consequence, we will first run a ‘pre-FDR’ exercise to prepare various tasks, and then the first FDR will run into the first CCRC exercise. The planning of the next FDR and CCRC (and indeed the cosmic data running planned at about the same time) is currently under review.

6. Conclusions
The ATLAS computing model as described in the Computing Technical Design Report is standing up rather well to the various exercises and commissioning tests performed so far. In addition, reviews of the proposed analysis model and comparisons with those of other data taking and LHC experiments have not exposed major unexpected issues. However, there are several major components still to be undertaken: the rapid calibration, the reprocessing at the Tier 1s and the group-based analysis and DPD production being amongst the most notable.

The issue of quotas remains significant in ATLAS. There are some adjustments to the model to compensate for this, but a full solution is required soon. This is an area where there is a risk of divergent solutions being found within the experiments, increasing the long-term maintenance effort for the experiments and for the resource providers (as has already happened in the absence of common high-level data management tools).

The current tests are adding the last required activities to the menu of Grid operations. In these tests, the growing involvement of physicist-users is invaluable in debugging. These early adopters suffer some hardship in their work, but are making a great contribution towards a successful outcome for the experiment.

7. References
[1] G Duckek (ed.) et al 2005 ATLAS Computing Technical Design Report CERN-LHCC-2005-002 ISBN 92-9083-250-9
[2] Drohan, J, 2007 The Atlantis event visualization program for the ATLAS experiment These proceedings
[3] R Alfieri, R Cacchini, V Ciaschini, L dell’Agnello, A Frohner, A Gianoli, K Lorentey, F Spataro 2003 VOMS, an Authorization System for Virtual Organizations 1st European Across Grids Conference, Santiago de Compostela, February 13-14, 2003
[4] F Rademakers and R Brun, 1998 Linux Journal Issue 51 (Metro Link Inc.)