Draft 0.6 July 31, 2007

ATLAS Requirements for the National Analysis Facility

The German ATLAS Groups

Abstract

The setup and operation of a National Analysis Facility (NAF) is needed as a cornerstone for a common and coordinated physics analysis effort of the 15 German groups involved in the ATLAS experiment at LHC. The NAF should emphasize services for the final stage of the physics analysis chain, such as interactive access, PROOF clusters, etc. In addition it should complement the existing Tier-1 and Tier-2 facilities for large scale batch processing via the Grid. The setup and services of the NAF should be kept flexible in order to quickly adapt to changing requirements, in particular for the startup phase. The document gives a brief overview on physics analysis in ATLAS and presents the basic requirements for the NAF.

1 Overview on ATLAS Offline Computing

The NAF is an important component of the ATLAS offline computing facilities available for the German groups and should provide complementary services to the Tier-1 and Tier-2 facilities. The main focus of the NAF are analysis-group activities and the support of individual users. In the following the ATLAS analysis model is briefly introduced.

1.1 ATLAS Reconstruction and Analysis Chain

The offline reconstruction and analysis of ATLAS data proceeds in several distinguished steps, as illustrated in Fig. 1. The first step is an Athena job for the reconstruction of the raw data (RAWD), which comes from the ATLAS Online/DAQ systems or from the simulation. The output of the reconstruction are the so-called ESD and AOD data sets. The ESD still contains detailed informations on entities such as the individual hits forming a track and the calorimeter cells. The AOD provides a more condensed summary of physical objects such as identified particles, combined tracks, jets in the calorimeter, etc. Reconstruction is in general performed in an ATLAS-wide organised manner. Individual users do not run a full reconstruction based on RAWD, except for development, debugging or testing purposes done by few experts.

The next step in the analysis chain are typically Athena jobs using AOD data as input which are performed either in an organized manner by physics groups or sub-groups or by individual users. For certain cases the information in the AOD might not be sufficient for the analysis and access to the ESD is required. Large scale access to the full ESD data sets is restricted to physics groups for organized and scheduled analysis and as such is not foreseen for individual users. The output of an AOD/ESD job is typically a ROOT-ntuple, either in a common ATLAS format (DPD) or a private ntuple.

The final step of the analysis is typically a ROOT job analyzing the ntuples and producing ROOT histograms as output. Large scale processing of RAWD, ESD and AOD data sets will be performed on the Grid, making use of the designated resources world-wide. The processing of RAWD (and presumably ESD as well) will get organized as a central task and operated by a shift team. For individual users' AOD analysis the Ganga [4] distributed analysis tool is developed in ATLAS which simplifies the usage of the Grid and the setup of ATLAS jobs. Ganga will take care to match jobs submission and data location, it provides job-splitting and submission and tools for job control and monitoring.

The requirements for the DPD/Ntuple analysis vary largely. In some cases local desktops or laptops may be sufficient, while others involve CPU or data intensive processing steps which are better performed in Grid batch jobs (directly or with Ganga) or on dedicated PROOF clusters.

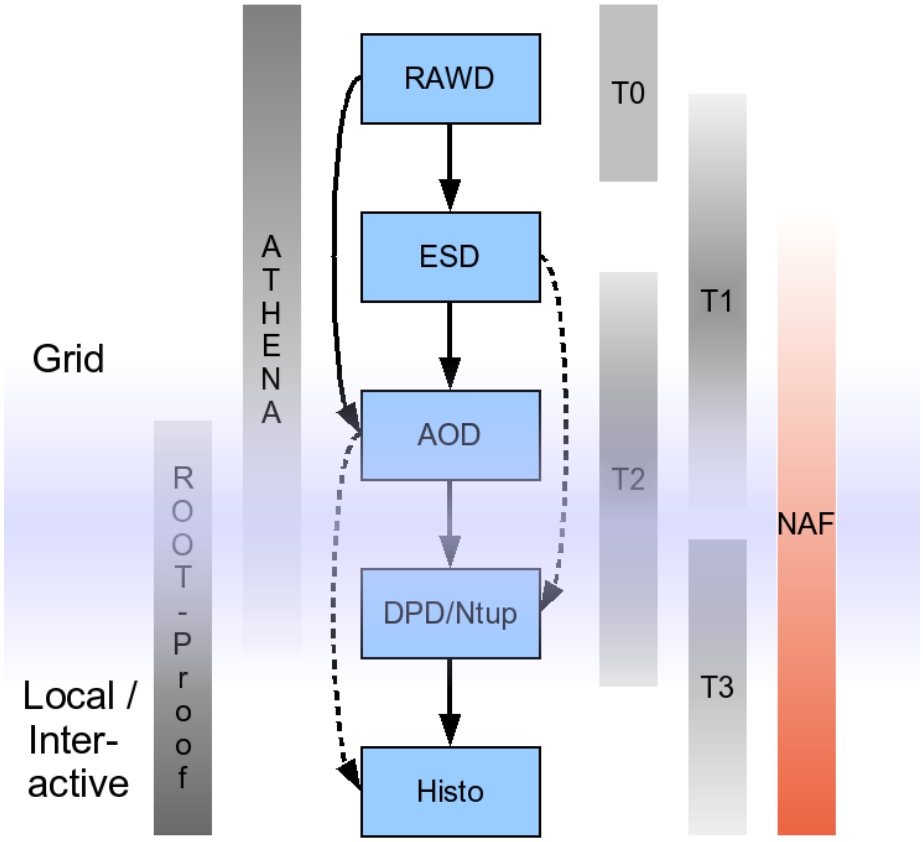

Figure 1: The ATLAS analysis chain and Tier structure

1.2 ATLAS Simulation

The official ATLAS simulation is performed at the Tier-1 and Tier-2 sites using a central production system which is operated by a shift crew. Substantial resources are foreseen in the Computing Model for simulation, additional resources such as Tier-3 or non-ATLAS sites may contribute on an ad-hoc basis. The ATLAS physics groups fill requests for the production of samples according to their needs, which will subsequently be processed by the production team according to the priorities defined by the ATLAS physics coordination. It can be expected that there is a substantial need for Monte Carlo production by physics groups and individual users in addition to the officially produced samples. Additional resources such as Tier-3, NAF, local farms or non-ATLAS sites have to provide this service.

1.3 Roles of the different Tier-s

The roles and requirements for the Tier-0, Tier-1 and Tier-2 centers are described in detail in the ATLAS Computing TDR [2]. The services and resources provided by these sites form the WLCG (Worldwide LHC Computing Grid) and are specified in the WLCG Memorandum-of-Understanding [3].

- Tier-0: Raw data storage; first calibration and reconstruction; distribution of derived data sets; very large storage capacity; located at CERN.
- Tier-1: Further calibration and reconstruction passes; storage of a fraction of raw and reconstructed data (ESD); facilities for organized reconstruction and analysis; large storage capacity; associated support for Tier-2 (and lower levels); 10 sites planned for ATLAS; German Tier-1 is the GridKa facility at FZ Karlsruhe.
- Tier-2: Central simulation and user analysis; storage of a fraction of summary data (AOD); typically 3 Tier-2 associated to one Tier-1 center; about 30 centres planned for ATLAS, with some 25 active users on average.
- Tier-3: Ntuple analysis; development, visualization.

The right part of Figure 1 illustrates the region in the reconstruction and analysis chain, which the different Tiers cover: Tier-0 exclusively serve first-pass reconstruction and calibration, the Tier-1 mainly serve central RAWD reprocessing and group-level analysis of ESDs and the Tier-2 provide access to the AOD data for individual users and storage capacity for the produced DPD/ntuples. The DPD/ntuple analysis proceeds at the Tier-2 and Tier-3 sites. The NAF combines and extends Tier-2 and Tier-3 services.

2 Role and Requirements for the NAF

The German ATLAS community encompasses 15 university groups and institutes with rather varying size and involvement in ATLAS computing. The setup of a NAF in Germany is an important premise to build a common basis and environment for a coordinated physics analysis of the German groups.

The services specified in the ATLAS Computing Model and the WLCG MoU focus on the production requirements and large scale batch processing in the Grid. Additional services and resources are required to support the physcis analysis user and interactive tasks.

The NAF will also act as basis for the "Virtual IT Centre" (VITC). The NAF, the Tier-2 and Tier-3 facilities within the Helmholtz Alliance should eventually be integrated into the VITC.
Basic NAF services are listed in the following:

Interactive Login for every member of the German ATLAS groups. The account should provide workspace on a shared filesystem and access to the currently used ATLAS software releases for software development, tests and debugging. Furthermore it should offer up-to-date user-interfaces to the Grid for jobs submission and data handling as well as a well maintained Ganga installation for access to the ATLAS distributed analysis. An AFS client installation should be made available in order to allow access to the CERN-AFS cell, which provides all the ATLAS releases and nightly builds, as well as group areas and the CERN-AFS home-directories.

High capacity storage space for users and groups should be foreseen to store group or private ntuples. The storage must be accessible via the Grid (gridftp, SRM, FTS) and locally (dcap, xrootd) and should include a back-up service. Depending on the access pattern, some part of the storage should be provided on high-bandwidth systems, e.g. DPD/Ntuples for PROOF analysis.

Full set of the ATLAS AOD locally accessible from the NAF, in conjunction with the Desy Tier-2 service. Having all AODs available on a single site largely simplifies interactive tasks such as development and debugging as well as event display and visualization.

Local batch queue for short jobs ($\approx$1 hour) for testing purposes.

Large-scale batch capacity should be provided via Grid-submission. The batch compute-nodes should be integrated into the Tier-2 setup. An ATLAS-Germany specific access to the batch resources will be ensured by a corresponding group in the Grid VOMS system.

Dedicated PROOF cluster as a subset of the NAF farm. PROOF is a promising tool for high throughput interactive analysis of data with root which is based on a sophisticated parallelization of the data and task flow. For lightweight applications and standard root files it is a proven concept, for ATLAS data it remains to be seen whether PROOF can be used beyond standard ntuples, e.g. AODs.

Full TAG database: ATLAS TAGS provide "micro-DST" information for each event and allow a fast pre-selection. Storing TAGs in a powerful database allows users sophisticated database queries, e.g. in order to test and optimize analysis scenarios.

Access to ESD should be provided for development and debugging. In the start-up phase many analyses might need large-scale access to ESD data sets, i.e. it might be more useful to provide as many ESD data sets as possible rather than the full AOD sample. For large-scale ESD access a dedicated service for the Conditions database might be required at the NAF.

Uniform NAF access: In case the NAF services are distributed over several physical sites, the details of service locations must be transparent for the users.

The ATLAS computing model provides in principle a rather detailed structure of services, distribution of tasks and resource allocation. However, several of the underlying assumptions have still large uncertainties, such as events sizes, processing times, simulation fraction, access modes, etc. Furthermore, it addresses steady-state operation, a solid model of the startup phase is impossible and a high degree of flexibilty will be crucial in the first months. Therefore it is important to keep setup and services of the NAF open for change in order to quickly adapt to evolving requirements, such as large ESD data sets in the start-up phase or extended PROOF capacity.

Further discussions and planning on the particular requirements of the start-up phase are in progress both in ATLAS as a whole and the ATLAS-D community. More concrete scenarios should become available in the next months.

References

[1] The ATLAS Computing Model, ATL-SOFT-2004-007, V1.2, 10 January 2005

[2] ATLAS Computing TDR, CERN-LHCC-2005-022, ATLAS-TRD-017, Jun 2005, `http://cdsweb.cern.ch/search.py?recid=837738`

[3] WLCG Memorandum of Understanding, CERN-C-RRB-2005-01/Rev., 15 June 2007 `http://lcg.web.cern.ch/LCG/C-RRB/MoU/WLCGMoU.pdf`

[4] J.Elmsheuser *et. al*, Distributed Analysis within the LHC computing, Proceedings of GES 2007, Baden Baden, May 2007, `http://edoc.mpg.de/316514`

A Resource Requirements 2007/8 – Version 0.1

Number of users: Most ATLAS-D members will need an account, i.e. 100-150 in total. Only a fraction will make regular and substantial use of the resources, assume about 50 such 'active users' and 10-20 simultaneous interactive logins for development, debugging or interactive analysis.

Shared filesystems: For home-directories (2 GB/user) and software installation (200 GB).

High capacity storage: Full AOD sample for nominal year corresponds to about 200 TB. For efficient direct event access and navigation the latest version should be available on disk.

In addition user storage should be provided on a disk-cache system with tape backend (T1D0-class storage). Capacity per 'active user' should be about 5 TB (this corresponds to 5×10^7 AOD or 5×10^6 ESD events).

Batch CPU capacity: The CPU resources needed for the analysis of the ESD/AOD/DPD samples are comparable with the requirements for a standard ATLAS Tier-2, i.e. about 500 kSI2k for a nominal LHC year. (Assuming 0.3 s/event for analysis this corresponds to processing 10^9 events per year and 'active user'.)

PROOF cluster capacity: About 20 CPU cores should be provided for an initial test setup, future evolution needs further discussion.

TAG database: TAG data of a nominal ATLAS year requires an Oracle RAC system of about 6 TB.