The National Analysis Facility at DESY - status and use cases by the participating experiments

S Aplin¹, W Ehrenfeld¹, A Haupt¹, Y Kemp¹, C Langenbruch², K Leffhalm¹, A Lucaci-Timoce¹ and H Stadie³

¹ Deutsches Elektronen-Synchrotron DESY, Hamburg and Zeuthen, Germany
² Physikalisches Institut, Ruprecht-Karls-Universität Heidelberg, Germany
³ Universität Hamburg, Hamburg, Germany

E-mail: steven.aplin@desy.de, wolfgang.ehrenfeld@desy.de, andreas.haupt@desy.de, yves.kemp@desy.de, clangenb@physi.uni-heidelberg.de, kai.leffhalm@desy.de, angela-isabela.lucaci-timoce@desy.de, hartmut.stadie@desy.de

Abstract. The German National Analysis Facility (NAF) was set up at DESY, starting end of 2007 in the context of the Helmholtz Alliance “Physics at the Terascale”. The NAF complements the DESY and the German Grid resources, and hence offers users from the German HEP institutes the best possible environment for data analysis. In the first part, the key aspects and components of the NAF are briefly presented with an emphasis on recent improvements. In the second part, the use cases of the three participating LHC experiments, ATLAS, CMS and LHCb, will be presented. Differences and commonalities in the usage of the NAF will be shown. Special emphasis will be placed on the usage of PROOF, whose usage on the NAF has been pioneered by CMS. It is now adapted by ATLAS. The third part will concentrate on how the NAF was used for detector optimisation studies in the preparation of one of the ILC LoI’s (ILD 2009), as well as how CALICE uses the NAF for the analysis of their data taken in several test beam experiments performed for detector R & D. Finally, future developments of the NAF are presented.

1. Key Aspects and Components
The main components of the NAF are shown in figure 1 including the amount of available resources as of October 2010. The main parts are:

- Additional hardware (CPU and storage) to the Grid sites
- Interactive work group servers
- A “local” batch system available from the work group servers
- A clustered file system for fast local access

As shown in Figure 1 all parts can access the Grid storage system (“dCache”) to read and write data. An additional clustered storage optimized for single client performance (“Lustre”) is available on the work group servers and the batch system. These machines are connected to the home directories provided for all NAF users in AFS. One important AFS feature is the global availability outside the NAF. The additional Grid CPUs are set up in a way to give the German users a dedicated share and higher priority compared to standard user jobs. The hardware is distributed over two DESY locations Hamburg and Zeuthen.
Figure 1: NAF schematic setup with amount of available resources as of October 2010.

2. Recent Improvements
Several aspects of the NAF design have been optimized according to the users requests. Hardware can be added for dedicated usage, CPU resources (work group servers, batch system) as well as storage (AFS/dCache/Lustre). To coordinate these resources, subgroups below the VOs can be introduced by the VO admins. They can also add and remove users to/from the subgroups. Restricted access of the dedicated storage resources can be guaranteed via quotas and POSIX rights set for individual subgroups for Lustre. AFS provides quota and access control lists (ACL) on a directory basis. The batch system provides the possibility to setup priorities for different subgroups. These can only be changed by the NAF operators and have to be setup once. Then priorities of jobs/users can be handled by VO admins.

To be aware of the experiments and user requests close work with the NAF User Committee (NUC) is necessary. Input to configuration of the batch system is given there and a cooperation to integrate PROOF are some examples for increasing the usability. In general, the changing and additional user requests are discussed on a technical level and the setup is subsequently adapted. Especially, the optimisation of new resources and the balance between storage and CPU resources and the needed data throughput are discussed.

3. ATLAS Germany NAF Setup and Usage
The NAF provides additional hardware resources to the ATLAS Germany community to boost their data analysis activities. For fast and efficient data analysis not only optimised hardware and infrastructure but also software is necessary [1].

To allow for a broad range of ATLAS analyses with up-to-date tools, ATLAS specific software is installed in addition to the centrally provided common software. The deployment of ATLAS software and tools, ranging from a wide selection of ATLAS releases and caches to flat ATLAS condition data taken from the DESY Tier 2 (T2), is performed in a (semi)automatic way and no individual user setup is necessary. Further, Grid tools, such as Panda and Ganga clients for distributed data analysis and DQ2 tools for distributed data management, are deployed and can be accessed directly or using simple and automated setup scripts.

The local batch system is mainly aimed at ntuple-based data analysis and hence optimised for fast turn around of short running jobs, i.e. with a few hours run time. For high profile analyses,
e.g., close to publication, a high priority batch queue is provided to ensure high throughput even with concurrent ATLAS analysis. Further, PROOF based data analysis is supported at the NAF. For more details see section 4.

The additional NAF dCache storage is used in different ways. Around 200 TB dCache space is added to the pledged ATLAS T2 storage to host the full AOD dataset. Therefore, the German ATLAS users can directly profit from the dedicated Grid CPU resources or the NAF batch system. 140 TB dCache space is assigned for user and group storage. User store their own or group ntuples for local access. Some part of the storage is used for archival of older Monte Carlo, which is needed by the German ATLAS community.

The major activities on the NAF batch system can be classified into three groups: data analysis, Monte Carlo production and fitting, e.g., Monte Carlo tuning and model fitting. 2010 is the first year with significant pp collision data and many user moved from Monte Carlo studies to data analysis. In the following some typical steps within the analysis workflow are discussed in more detail.

**ntuple production:** As the NAF hosts all AODs ntuple production can be done on the T2 Grid cluster or on the NAF batch system profiting from the already discussed benefits.

**ntuple import:** Within ATLAS some physics groups produce centrally dedicated ntuples. These are imported to the NAF storage (dCache or Lustre) for further use. So far no clear advantage of one storage system has been seen. Data import to dCache is much easier as it is fully integrated into the ATLAS distributed data management.

**ntuple analysis:** The main aim of the NAF is the support of data analysis. Ntuple analysis is mainly performed on the NAF batch system as it has advantages over the Grid. Besides the already discussed benefits it allows for easy access to local files, e.g., software. Output is usually written to Lustre. If the used analysis framework supports PROOF, PROOF on the batch system is used. If the dataset is small enough due to previous skimming steps on the Grid or the local batch system, some user perform their final analysis on a single work group server.

Summarising, the NAF is used by the ATLAS Germany community. More and more user work on data analysis. ATLAS NAF users contributed to a large number of conference notes for summer 2010. Some NAF users significantly contributed to early ATLAS public results, e.g., references [2, 3].

Much more data is expected from the LHC in 2011 than in 2010 increasing significantly the data volume. It is expected that one bottleneck is the needed I/O rate between storage and worker nodes. In order to increase the current rate the infrastructure can be improved, e.g., upgrading the network connection to 10 GE links or Infiniband, and the throughput from the storage can be increased. It was also observed that a non-optimal file distribution on the dCache system can degrade overall bandwidth significantly.

4. **PROOF on the Batch System**

The “parallel ROOT facility” (PROOF) from the ROOT framework provides a mechanism to distribute the load of interactive and non-interactive ROOT sessions on a set of worker nodes optimising the overall execution time. While PROOF is designed to work on a dedicated PROOF cluster, the benefits of PROOF can also be used on top of another batch cluster with the help of temporary per user PROOF clusters.

Within the NAF ATLAS and CMS jointly developed a lightweight tool which starts a temporary PROOF cluster on the NAF batch cluster [4]. Every user can setup his own PROOF cluster on demand with the resources necessary for his work. To execute the task of configuring, starting and stopping a temporary PROOF cluster in a transparent way, a python script was written, based on some previous work for CMS.
PROOF at the NAF is currently mainly used by ATLAS. The SFrame package \cite{5} is very popular. It is a ROOT-based analysis framework, which transparently implements multi-processor usage via PROOF or PROOF Lite. One direct advantage is the optimised assignment of data to worker nodes. PROOF has build-in procedures which minimises the overall turn-around time.

The current usage of PROOF is small, less than 3\% of the total batch jobs are in PROOF sessions in terms of wall clock time. As most of the PROOF sessions are started interactively, they are idle most of the time waiting for user requests. When processing data the CPU efficiency can reach 100\%. The average PROOF efficiency in terms of CPU time is less than 1\%. This low efficiency needs to be addressed if PROOF usage increases. For example, by terminating the PROOF session after a specified, idle interval or increasing the overcommitment of jobs to CPU slots. Data analysis from users and dedicated tests have shown that the dCache storage in principal can stream data to sufficient worker nodes without problems. Bottlenecks can arise from limited network band width or asymmetric file distribution over dCache pools.

5. CMS NAF Setup and Usage

The additional computing and storage resources at the NAF are used by the German CMS groups for a variety of tasks. The main use cases are the following:

- (prompt) data analysis, requiring fast turn around of jobs.
- production of special MC samples
- development of analysis tools (e.g. kinematic fits)
- calibration and alignment

To cover these use cases, the NAF environment has been adapted to the needs of the CMS experiment. All German CMS groups contribute to this effort. The CMS software, CMSSW, is installed on AFS and all releases needed by the individual physicists are installed. Hence, also releases that are not centrally installed by CMS at Grid sites are available. All files hosted at the DESY CMS-Tier-2 site are accessible from the interactive nodes and the worker nodes of the "local" batch system. The CMS remote analysis builder, CRAB, and the grid-control toolkit have been modified to support submission to the SGE batch system used at the NAF. This allows the users to switch transparently between submission to the Grid and to the "local" batch farm. A German CMS support team consisting of members from every German group has been established to assist the users and maintain the documentation.

The additional Grid storage resources are used to offer large user directories and host additional data sets. In the CMS computing model the output of an analysis job on the Grid is copied back to a user directory at the "home" site. The large home area allows the users to have these output files at the NAF for their further analysis.

The data volumes from the 2010 running of the CERN LHC are still relatively small and complete primary data sets can be stored on the additional Grid storage. Therefore, all analysis steps can be performed on the NAF. These additional data sets amount to 160 TB and include all collision events triggered by electrons, photons, muons, or jets and the most important signal and background Monte Carlo samples. Figure 2 shows how many different users accessed these data sets in the period of 60 days. Many data sets have more than 20 different users while only a small fraction has not been accessed in this period. Hence, the selection of interesting data sets works very well.

The transfer of new data to the NAF works very well and new runs are available at the NAF within a day after data taking. This quick availability of data sets and the hosting of all interesting Monte Carlo samples resulted in significant contributions to published results and even whole analyses performed on the NAF.
Figure 2: How many TB of data is used by how many distinct users? From 160TB of data stored at DESY 20TB are in fact accessed by more than 20 distinct users.

6. LHCb NAF Setup and Usage
Centralized production of data summary tapes for the LHCb collaboration is performed on the Tier 0 and Tier 1 centres which are accessible via the Grid tools, as described in the LHCb computing model [6]. The NAF’s resources allow for a fast turn-around time, important during development and testing of users’ physics analysis code, and accompanying ntuple production. In this way, those resources can be employed very effectively by LHCb users. Tasks performed on the NAF are analysis code development and ntuple production as well as computation intensive data fits.

The LHCb NAF support provides a web interface which can be used to download and install LHCb software and data. With the fast turn-around time, development and testing of users’ physics analysis code and accompanying ntuple production is very efficient at the NAF. Various user-analyses have profited from these capabilities, one example being the analysis of the lepton-flavor violating decay $\tau \rightarrow \mu\mu\mu$.

The study of CP-violation in the B sector requires computation intensive unbinned Maximum Likelihood fits. One prime example is the extraction of the CP violating phase $\phi_s$ using the decay $B_s \rightarrow J/\psi\phi$, which requires a multidimensional fit with up to 30 parameters. Large numbers of jobs are needed to validate the fitting procedure and to extract expected parameter errors. A typical study consist of 1000 jobs requiring a CPU time of 14 minutes each. For small amounts of data a special statistical method proposed by Feldman and Cousins [7] is used to get correct coverage. This method is extremely computation intensive. Typically at least 200 jobs each consuming about 4 hours of computation time are needed. Due to the large computing capacity of the NAF results for these extensive studies can usually be obtained within one day.

7. Example of ILC NAF Usage
The ILD letter of intend required studies of the impact of machine background on track reconstruction efficiency. The global tracking algorithm used for physics studies needed adapting to cope with the huge increase in the number of background hits, $O(10^5)$ in the vertex detector. This was achieved by using the short batch queue with $< 1$ hour wallclock time (to have fast turn around time) and a large number $O(100)$s of concurrent jobs. The Lustre file system was used to store the large background data files with $O(100)$ GB. The software was made available locally in AFS. This way it was easy to recompile the code and resubmit the jobs.

The fast turn-around time achieved made for very efficient prototyping, almost in real-time. Importantly, due to the limited time available for the analysis, this could be set up with relative ease, using a small number of python scripts to submit and monitor the jobs.
8. Example of CALICE NAF Usage
Examples of CALICE’s usage of the NAF include the GEANT4 validation with AHCAL data: Track multiplicity and custom Monte Carlo generation (An example plot in Fig. 4).
These were done using the NAF because of the ability to work with scripts in a homogeneous environment which provided scope for large parallelisation. At the same time it was possible to retain efficient access to Grid storage.

9. Summary
The use pattern changed from the pre LHC area of MC studies to the LHC data taking area with early data analyses and performance studies. The NAF in communication with the experiments in form of the NUC is following these changes to adapt the NAF design accordingly. Further, user requests are passed on via the NUC to the NAF operators. Many user requests are solved this way.

All participating VOs have established the full analysis chain at the NAF. The NAF is used by the VOs in addition to the global Grid and local resources available at the institutes.
The use pattern will continue to evolve as the LHC will approach its design luminosity. It is therefore important to keep up with the developments to keep the NAF well equipped for changes in analysis work flows. More fine tuning might be useful to introduce locality in now over two sites distributed facility, especially to achieve short latencies from batch nodes to storage. To achieve this, close collaboration between the VOs and the NAF operators via the NUC is mandatory. There is always work in progress.

Some of the used software solutions might not have a clear support in the future like the Lustre cluster file system and the SGE batch system. This requires the NAF operators to look for adequate substitutes to ensure the future usability of the National Analysis Facility.

Reference