The NAF: National Analysis Facility at DESY.

Andreas Haupt and Yves Kemp
DESY, Notkestr. 85, 22607 Hamburg
E-mail: andreas.haupt@desy.de, yves.kemp@desy.de

Abstract. Within the framework of a broad collaboration among German particle physicists, the strategic Helmholtz Alliance “Physics at the Terascale”, an analysis facility has been set up at DESY. The facility is intended to provide the best possible analysis infrastructure for researches of the ATLAS, CMS, LHCb and ILC experiments and also for theory researchers. In a first part of the contribution, we will present the concept of the NAF and its place in the existing distributed grid landscape of the experiments. In a second part, the building blocks of the NAF will be detailed with an emphasis on technical implementations of some parts:
- Usage of VOMS for separating grid resources between collaboration-wide and NAF-specific resources.
- interactive and batch cluster and integration with PROOF.
- usage of grid proxies to access work group servers and AFS.
- the usage and operation of Lustre for fast data access.
A special focus is the seamless integration of the facility into the two geographically separated DESY sites and its implications. In a third part, the experience of running the facility for one year will be reported.

1. Introduction
In 2007, the Strategic Helmholtz Alliance ’Physics at the Terascale’ was founded, a structured research network comprising 17 universities, 2 Helmholtz institutes and 1 Max Planck Institute. The Alliance acts as a tool for a more effective collaboration, in particular between experimentalists and theorists. The Alliance covers four research topics, addressing the fundamental questions of particle physics, novel detector development, accelerator science and distributed computing.

One of the work packages in the ’distributed computing’ project is the NAF: National Analysis Facility. This facility gives Alliance users additional and complementary resources for their analysis. The needs are estimated to be 1.5 of an average LHC Tier-2 center, with an emphasis on data storage. The proposal of the NAF foresees a potentially distributed NAF.

The choice of DESY can be explained with its role in the computing models of the ATLAS and CMS experiments which foresee user analysis at the Tier-2 centers [1]. The DESY Grid infrastructure is used by the ATLAS, CMS and LHCb experiments as a Tier-2 center, so it is natural to place the NAF at DESY. ILC does not yet have a detailed computing model, however, DESY plays an important role in ILC computing, which also makes DESY a natural candidate for an ILC analysis facility.

1 on behalf of the NAF team at DESY
2. NAF layout and building blocks

In 2007, discussions with the experiment groups began on what services and resources the NAF should contain. The following key points summarize the results:

- Interactive access to large scale computing resources coupled to experiments’ data.
- All data at hand for analysis, no need for data moving.
- New tools, like parallel analysis frameworks or powerful databases.
- Seamless integration of grid authentication and authorization methods with local systems.
- Additional grid resources, dedicated to people working in the Terascale Alliance for batch processing.
- Good support and documentation.

Experts from the DESY IT and DV divisions devised a scheme which consisted of the building blocks depicted in fig. 1.

In the following the different components and their interaction will be briefly presented. Some aspects will be described in more detail in later sections. The last section summarizes the resources.

2.1. Grid: dCache storage and cluster

Well established and already in production use since 2004 is the DESY Grid, a generic infrastructure for many different virtual organizations and projects. In the LHC context, it serves as a Tier-2 for the ATLAS, CMS and LHCb experiments. The ILC VO is fully supported by the DESY Grid infrastructure, all core servers are run by DESY. These grid resources are integrated into the work flows of the experiments, which have chosen to use the grid as their workhorse. It is clear that the grid dCache storage is the main data import, export and exchange system for data. It should be accessible identically from all systems in the NAF. Together with the experiments, the decision was made to use the existing storage elements, and just to equip them with more space to hold additional data. A first consequence of this is that normal grid users (e.g., not part of the Alliance) can also access this data. Special areas in the SE can however be configured such that only Alliance users have write access to them. The tools used
to access the data are the usual tools of the collaboration.

The Grid Cluster was extended to offer Alliance users additional resources for batch processing. All computers in the Grid Cluster are configured identically and can be used by all users. The underlying batch system was configured to offer Alliance users a special fair share, independent of other groups. They also have a higher priority, i.e., in case of a full cluster, their analysis jobs will be prioritized over other jobs in the limit of the fair share. One important prerequisite is the introduction of VOMS groups to distinguish Alliance users from other users. All VOs using the NAF have created such a group, all users have to do prior to job submission is create a VOMS proxy with the adequate extension by typing e.g. `voms-proxy-init --voms atlas:/atlas/de`.

The submission methods are otherwise completely identical to the collaboration tools, e.g., GANGA or CRAB.

2.2. Interactive, Local Batch and PROOF

During an analysis, some parts of the work are of a nature unsuitable for the grid. For tasks with an usage scheme orthogonal to a grid one, additional complementary computational resources are made available in an interactive cluster with an attached batch system with local submission. Users log into the work group servers in a first step. The work group servers are intended as developer and short testing environments. For longer jobs, they can use the local batch farm. The Grid Cluster is however recommended for traditional batch processing.

As resource manager the highly configurable and mature Sun Gridengine (SGE) [5] has been chosen. It has been setup to prefer short running jobs (< 1 hour) but also allows for jobs running up to one week. Resource allocation to the different experiments is done via a fair share that respects current as well as historical usage.

All batch jobs are provided with an AFS token so that they can access, e.g., the user’s home directory.

In the last years, PROOF, the Parallel ROOT Facility, was developed, and users want to use this for a fast interactive analysis of their data. A priori, a setup of such a facility contrasts with a traditional batch facility. However, using SGE, the NAF can integrate both: By the use of so called parallel environments a user can reserve many CPU slots spread over many machines at once [2, 3]. They will be used to run PROOF-workers on them. This allows for a proper accounting and resource usage monitoring.

2.3. AFS and Parallel Cluster FS

From the interactive and local cluster machines in the NAF, not only the dCache storage element is available, but also two additional spaces: An AFS based space and a parallel cluster file system implemented using Lustre.

AFS is used in the NAF to host the users home directories as well as group directories. The latter are often used by the experiments to install specific software. AFS allows for remote access, the next paragraph will detail how the authentication and authorization is done for the NAF AFS cell.

The cluster file system Lustre fulfills a different role: It is intended to serve as a fast scratch space that contains data which is analyzed very often and at a very high throughput rate. The Lustre file systems can only be accessed from the interactive and local cluster machines of the NAF. The NAF operators opted against an opening via SRM when designing the system, as no products were available that allowed a secure and efficient read and write access both locally and via the remote protocol with the current version of Lustre (1.6.x). It must be kept in mind that the NAF supports among others both ATLAS and CMS, which have very strict requirements on the accessibility to data.
Figure 2. CMS analysis using PROOF, with different number of clients. Three access types to Lustre are compared: Local reading via InfiniBand (upper blue curve), WAN connection using unmodified TCP buffer setting (lower red curve), and WAN connection using optimized TCP buffer settings (middle green curve) [2, 3].

2.4. Authentication in the NAF: Kerberos/X509 integration

Users do not need to remember yet another password when logging into the NAF. Access is granted via proxy certificates that are already widely used in the grid world. To allow transparent access to the user’s AFS home directory right after the login a Kerberos5-X509 integration has been set up. This is done using the Kerberos5 implementation Heimdal [4]. It allows for an authentication at a Kerberos5 realm using Globus proxy certificates. Upon login via gsssh a self-written PAM module cares for the automatic generation of a Kerberos ticket and an AFS token.

Using a Kerberos-enabled MyProxy service, also the opposite is possible: Generate a proxy certificate from a Kerberos ticket. This is currently in beta-testing phase, and will allow users to always have a valid proxy certificate, potentially with a VOMS extension. This will ease user’s work as they will be authenticated and authorized in an homogeneous way.

2.5. Interconnect

The Lustre file system is connected to local machines via InfiniBand (IB DDR, double data rate, 20 Gbit/s at at latency of 2 µs). As DESY has two sites, this InfiniBand connection can not be used for nodes at the remote site. Instead, a normal TCP/IP connection is used here, using the dedicated 10 Gbit link between the two sites. To cope with the latency, the TCP buffer size has been adapted on the SL4 nodes. In fig. 2, the same CMS analysis using PROOF was performed using Lustre over local InfiniBand, using default TCP setting over WAN, and using the modified TCP settings over WAN. One can clearly see the improvements in speed.

3. Running experience

All parts of the NAF are in place and actively used. Around 300 users have registered for the NAF, the majority from institutes outside of DESY, which makes clear that DESY is offering a service to the national particle physics community. In retrospect, this also reinforces our decision to not build on the existing DESY infrastructure for interactive computing, but to build something new, with the only legacy components being the DESY Grid clusters and SE (which are anyway already a service to the global community).
3.1. Resource overview

As of April 2009, the approximate resource allocation is the following:

- Grid cluster: 2500 CPU cores, of which 800 are NAF shares.
- Interactive, local batch (with PROOF): 900 CPU cores.
- dCache: 1300 TB, of which 800 TB are pledged to the experiments by an existing MoU (e.g. Tier-2). The rest (500 TB) constitutes the NAF part. Additional 400 TB are in preparation.
- Lustre: 100 TB

3.2. Resource utilization

The different subsystems play different roles in the analysis steps. It is therefore impossible to give one single measure of the overall utilization. Two plots are provided in fig. 3 as separate examples for the grid and local batch parts. Over time, the average system utilization is around 80%, which is very good when one aims to maintain responsiveness to significantly fluctuating demands. Not easily measurable, but a very important factor for the user acceptance is the fast or immediate availability of resources for short and interactive applications. Together with users, the setup of the batch system is constantly evaluated and tuned to achieve this goal, potentially at the detriment of an overall cluster efficiency.
3.3. Support and documentation
The users’ acceptance is fostered by the good support and documentation in and around the NAF. The NAF operators provide and maintain documentation about the basic functionalities of the NAF. In addition to the expert level support, they also provide a generic first level support for all fabric related issues.
The experiments also have organized their own support and documentation for specific issues relevant to them. In 2008, several training sessions were organized by the experiments and supported by the NAF operators to help new users getting started on the NAF. These were very well received by the community.

3.4. NAF Users Committee
A good communication between NAF operators and NAF users is important to ensure the best possible handling of problems and an optimal planning of future changes or additions to the NAF, but also to canalize discussions between users on these issues.
The NAF Users Committee (NUC) has been created, in which each experiment as well as the operators are represented by two members. In the monthly meeting of the NUC, current and future issues are discussed between the users, and the status of the NAF is being presented by the operators. The NUC is also the place where requests are exchanged between the representatives of the users and the operators.

4. Summary and outlook
The NAF is well established among the physicists of the Terascale Alliance. Both the grid part and the local part are used by many of them to carry out their work.
In the future, the NAF will be expanded by additional resources to cope with the demanding needs. New services might be implemented, such as the ATLAS TAG database which is currently being set up.
Fast access to data for analysis is the key point of the NAF. The access from the different computing parts to the different storage systems will be benchmarked and, if necessary, improved. This is only possible with standardized application benchmarks like the HammerCloud suite developed by ATLAS.
The NAF operators will closely follow the first real challenge: The start of LHC data taking. Together with the users, the NAF setup will be adapted to best match the experimentalist’s needs and offer them the best possible analysis facility.

5. Acknowledgments
The NAF is supported by the HGF Alliance “Physics at the Terascale”.

References
[1] ATLAS and CMS Technical Design Reports (CERN-LHCC-2005-022, CERN-LHCC-2005-023)
[2] Integrating interactive PROOF into a batch system W. Behrenhoff and Hartmut Stadie, CHEP09 proceedings
[3] Development of Interactive Analysis Tools for the CMS Experiment. Wolf Behrenhoff, diploma thesis, University of Hamburg 2008 http://www.desy.de/~vbehrenh/diplom.pdf.
[4] http://www.h5l.org/ Homepage of the Heimdal project.
[5] http://gridengine.sunsource.net/ Homepage of the SGE project