Evolution of Interactive Analysis Facilities: from NAF to NAF 2.0

Andreas Haupt, Yves Kemp, Friederike Nowak
DESY, Notkestr. 85, 22607 Hamburg
E-mail: friederike.nowak@desy.de

Abstract. In 2007, the National Analysis Facility (NAF) was set up within the framework of the Helmholtz Alliance “Physics at the Terascale”, and is located at DESY. Its purpose was and is the provision of an analysis infrastructure for up-to-date research in Germany, complementing the Grid by offering a interactive access to the data. It has been well received within the physics community, and has proven to be a highly successful concept.

We will review experiences with the original NAF, and discuss both the resulting motivation and constraints for the transition to an evolved model. We call this new facility the NAF 2.0. We will present a new setup including its building blocks and user handling, and give an overview of the current status. The integration of new communities has broadened the range of the analysis facility beyond its primary focus on LHC and ILC experiments. To finish, an outlook on further developments like the adoption of new technologies will be given.

1. Introduction
Since 2007, the National Analysis Facility (NAF) is operated at the two DESY sites in Hamburg and Zeuthen. Its purpose is to provide additional and specialized resources to members of institutes participating in the German Helmholtz initiative “Physics at the Terascale”[1] for their analyses. Initially it was aimed at physicists working in ATLAS, CMS, LHCb, ILC, and CALICE coming from around 20 German institutes.

During its operation, the NAF resources were extended up to a maximum of 4500 CPU cores, 600 TB of high-performance Lustre file space and 5 PB dCache space, in addition to Tier-2 MoU pledged space. The hardware and the systems were kept up to date, e.g. by upgrading the operating system when needed by the experiments. No essential changes were applied to the NAF as presented at CHEP 2009 [2]. It became clear, however, that a fundamental redesign was necessary.

In the following, the setup of the NAF as designed in 2007 and operated since then will be reviewed. From the lessons learned we derive design requirements to be implemented in the NAF 2.0.

2. The NAF Infrastructure
The NAF was set up with only minimal overlap with the “normal” DESY infrastructure (see figure 1). A dedicated installation and configuration management was introduced, and user registry, kerberos realm and AFS cell were completely decoupled. This choice was motivated by the uncertainty whether the NAF infrastructure would be extended beyond the two DESY sites.
Separating the NAF from the rest of the DESY infrastructure allowed a very fast setup and a steep increase in terms of installed resources and registered users. Some requirements could easily be satisfied in the NAF, but would have been very difficult to implement within the normal DESY infrastructure. The drawback of this choice is apparent: the NAF did not necessarily profit from advances in the DESY infrastructure and NAF users could not use additional resources and services. Also, the administration overhead became evident once larger clusters were set up for different groups within the DESY infrastructure. It is therefore clear that a redesign of the NAF needs methods on how to integrate large resources for a dedicated community into the general purpose DESY infrastructure.

**Figure 1.** Left: Infrastructure schema for the DESY sites and the original NAF. Right: The NAF 2.0 will be integrated into the DESY infrastructure (oversimplified view).

### 3. Centrality of Data

The NAF is intended to be the place where all data relevant to an analysis is stored and computation can be done in an efficient way. This led to the introduction of a cluster file system for fast analysis, connected to the compute nodes with a high-bandwidth, low-latency InfiniBand network. At the same time, due to the two-site setup, this storage was exported over a wide area network with limited bandwidth and high latency. Detailed monitoring of the worker nodes has shown that LHC analyses on the NAF do not make use of the high bandwidth offered by InfiniBand. Later, compute nodes without InfiniBand were purchased, allowing to acquire more compute nodes. However, the two-site setup remained a NAF concept.

In NAF 2.0, mounted file systems will not be accessible over WAN, thus enforcing the centrality of data for analysis. This is not in contradiction with the current movement towards federated access over WAN [3] or the CERN remote Tier-0 [4]. While these efforts are targeted to production-like tasks with a sequential access to data and little user interaction, the NAF hosts analyses which require random reads and substantial user interaction. These are directly affected by latency. Bandwidth is less of an issue as long as the line is not congested [5].

Centrality of data also means that NAF CPU resources are located in proximity to the Grid data. This is easily feasible for the CMS, LHCb and ILC experiments, because their Grid data is either located in Hamburg (CMS, ILC) or Zeuthen (LHCb). The case of ATLAS is more difficult, as for this experiment DESY hosts a federated Tier-2 and therefore provides storage capacity at both sites (see figure 2). A closer look at the storage shows that the usage is split. The “ATLASDATADISK” part is nearly exclusively accessed from Grid compute nodes, while the “ATLASLOCALGROUPDISK” part is nearly exclusively read from NAF compute nodes.
To enhance centrality of data for the ATLAS experiment, the “ATLASLOCALGROUPDISK” part of the Grid storage in Zeuthen will be reduced in size, while it will be enlarged in Hamburg. This also means that ATLAS analyses should be exclusively performed on the Hamburg part in NAF 2.0.

Figure 2. Simplified schema of Grid and NAF installations on the DESY sites, shown for LHC experiments. Except for ATLAS, for each experiment the data is located on only one site.

4. Proximity to the Grid vs. Proximity to “Users Laptop”
In a complex analysis chain, many steps are necessary, ranging from large scale data reduction steps down to changing the binning of the resulting final plot. The NAF has to position itself within this range. In 2007, users needed an infrastructure which was closely connected to the Grid not only for the data but also for analysis task workflows. The ideal NAF at that time would have had no boundary to the Grid. Derived from this criterion was the technical design decision to avoid the usage of traditional passwords for authentication, but to rely exclusively on X.509 Grid certificates for accessing the NAF via ssh or NAF AFS from outside.

This was very advanced in 2007, and our hope was that X.509 would find a broader way into operating systems and access tools and find adoption in other communities than the WLCG one. This, however, did not happen. Instead, more and more access tools came up that only relied on passwords. Changing the primary authentication method to passwords in NAF 2.0 became a prime design goal.

The initial workgroup-server concept foresaw a small application footprint, different from a traditional one. Desktop managers and tools that come with them were not installed. The reasons were twofold. One reason was to make the life of administrators easier, and the second was to prevent remote users from using tools that would not have been very usable anyway due to the high latency effects. Such applications, like graphical file browers, image viewers or graphical IDE should be run on the users’ laptops.

It has been shown that users increasingly need such applications on the same machine where they have access to the data. With tools like NoMachine “NX” [6], the penalty of low bandwidth and high latency networks became acceptable. Unfortunately, such tools were not usable as they do not offer X.509 authentication. For NAF 2.0, the workgroup-server needed to be equipped with more graphical applications, and offer access via “NX” or similar products.

The dCache [7] systems were set up before 2007 and hence without the NAF in mind. They were clearly integrated into the WLCG setup and the experiments data flows. Very little attention was put on local access to metadata and data by the experiments. This has changed. The importance of fast and reliable access to data was recognized by many providers of storage
systems. As one result, NFS v.4.1 and its feature pNFS was introduced. The dCache server implementation of NFS v.4.1 was tested in the last years and is now ready for usage. Actually, it is already in use at DESY for non-HEP purpose [8]. With the LHC experiments migrating to SL6, client can mount dCache just like any other mountable file system, with all the advantages for data and metadata access. Deploying NFS v4.1 and pNFS will be an important feature for NAF 2.0 SL6 machines.

5. Building Blocks of the NAF

The NAF features interactive workgroup-servers, a batch system for submission via the workgroup-servers and additional resources within the DESY Grid farm (see figure 3). All of these are standard two-socket X86(-64) systems with 2-4 GByte RAM per core and Gigabit Ethernet network. These building blocks are sufficient to fulfill the requirements of the experiments, with one exception: for running detector alignment software, a system with more cores and memory was needed, and could be integrated. During the setup phase, several options existed for implementing a PROOF cluster. We opted against a separate, dedicated PROOF cluster but rather developed an integration into the SGE batch scheduling system [9]. This was accepted quite well in the first years, but in later years users moved away towards Proof-On-Demand on the batch cluster (which is just a bunch of normal jobs from the point of view of the batch system) and Proof-Lite on the workgroup-server (which does not need a batch system) [10][11]. The additional Grid resources resulted in a dedicated resource share for German users. This was implemented using VOMS groups within the experiment VOs. This was well accepted in the first years of the setup. However, as experiments turned away from user submission towards pilot factories, the VOMS mechanism did not work anymore. Efforts within Germany to implement their own pilot factories were only partially successful. We decide to keep this building block as we know from experiment monitoring that many German users are using the DESY resources and profit from these additional resources. No new building blocks are required for NAF 2.0, just some adaptions as described above.

The NAF features AFS file service, a fast local cluster file service, and access to the Tier-2 dCache instance. This storage instance is provided with additional resources which host data relevant to analysts from German institutes. This setup proved to be the center stone of the whole NAF setup. As written above, dCache will be further integrated into the local NAF 2.0 setup when using NFS v4.1 as access protocol. AFS is well received by the community as a global file system and well known for home and group directories and will be continued.

The fast local cluster file service was introduced also in 2007. The idea was to keep 10% of the data located in dCache as copy in the cluster file system, and allow for storage of larger personal nTuples which in turn could be accessed with higher single-stream performance than the dCache system and higher capacity than the AFS file service. A fast local cluster file service is foreseen for NAF 2.0, but it will be observed how dCache access via NFS v4.1 as well as larger AFS user volumes can enhance performance and user experience. It could well be that the importance of the cluster file service will decrease if overtaken by dCache and AFS.

The interconnect between all other building blocks is currently a TCP/IP network. As described above, no InfiniBand network is currently in use. Several 40 GBit network constitute the network backbone of DESY-HH computing center.

6. NAF User Community and NAF Usage

The initial user communities were physicists from institutes participating in the Helmholtz Alliance "Physics at the Terascale". This included the experiments ATLAS, CMS and LHCb as well as ILC and CALICE. Plans to also offer resources for theorists from this alliance were never realized, partly due to low acceptance of Grid certificates in this community. With NAF 2.0, new
communities can join more easily both for technical and organizational reasons. A participating community does not need to rely on Grid certificates. On an organizational and political side, as NAF 2.0 is integrated into normal DESY infrastructure, all groups that can use normal DESY resources can in principle also use NAF 2.0 if a participation in resource provisioning is guaranteed. As first new communities, the HERA experiments H1, Zeus and Hermes, as well as BELLE 2 and the HERAFITTER project will use NAF 2.0. These communities are not restricted to German institutes. It will be important in NAF 2.0 to well document resource usage of the different projects.

The NAF Users Committee (NUC) has been set up to steer the technical implementation of the NAF. The NUC is constituted of delegates from the ATLAS, CMS, LHCb and ILC experiments as well as IT representatives. In NAF 2.0, also other communities might join the committee.

The NAF is well accepted in the German HEP community. Although situated at DESY, the usage fraction from DESY physicists only constitutes around one third. Many smaller German institutes rely on NAF resources for peak analyses, as they do not own larger private computing infrastructures (see figure 4).

Figure 3. Left: Building blocks of the original NAF. It consisted of workgroup-servers, a batch farm, extra resources within the DESY grid cluster, an own AFS cell, a fast cluster file system, and additional grid storage. Right: Changes in NAF 2.0 from the user perspective.

Figure 4. Resource usage by institutes for CPU (left) and fast cluster file system, represented by Lustre (right).
7. **Summary and outlook**
The NAF was and is very successful and important for physicists from German institutes. The successor, called NAF 2.0, is ready to be used, and migration has started. We expect to have completely moved all users by spring 2014. Evolution will continue, with the double benefit of always offering a modern facility to NAF users, and to provide developments done in the NAF 2.0 context also to DESY users from all communities.

**References**
[1] http://www.terascale.de/
[2] Kemp Y and Haupt A 2010 *J.Phys.Conf.Ser.* **219** 052007
[3] Bauerdick L, Benjamin D, Bloom K, Bockelman B, Bradley D *et al.* 2012 *J.Phys.Conf.Ser.* **396** 042009
[4] Andrade P, Bell T, van Eldik J, McCance G, Panzer-Steindel B *et al.* 2012 *J.Phys.Conf.Ser.* **396** 042002
[5] Fuhrmann P, Gaußhuber M, Kemp Y and Ozerov D 2012 *J.Phys.Conf.Ser.* **396** 042020
[6] https://www.nomachine.com
[7] https://www.dcache.org
[8] Millar P, Fuhrmann P, Mkrtchyan T and Schwank K these proceedings
[9] Behrenhoff W, Ehrenfeld W, Samson J and Stadie H 2011 *J.Phys.Conf.Ser.* **331** 072063
[10] http://pod.gsi.de/
[11] Malzacher P and Manafov A 2010 *J.Phys.Conf.Ser.* **219** 072009