Journal of Physics: Conference Series

OPEN ACCESS

The DESY Grid Centre

To cite this article: A Haupt *et al* 2012 *J. Phys.: Conf. Ser.* **396** 042026

View the article online for updates and enhancements.

Related content

- The NAF: National analysis facility at DESY
  Andreas Haupt and Yves Kemp
- A multi VO Grid infrastructure at DESY
  Andreas Gellrich
- Evolution of Interactive Analysis Facilities: from NAF to NAF 2.0
  Andreas Haupt, Yves Kemp and Friederike Nowak

IOP ebooks™

Bringing together innovative digital publishing with leading authors from the global scientific community.

Start exploring the collection–download the first chapter of every title for free.

This content was downloaded from IP address 131.169.253.70 on 02/02/2021 at 13:02

Journal of Physics: Conference Series **396** (2012) 042026 doi:10.1088/1742-6596/396/4/042026

The DESY Grid Centre

A Haupt[1], A Gellrich[1], Y Kemp[1], K Leffhalm[1], D Ozerov[1], P Wegner[1]

[1] Deutsches Elektronen-Synchrotron DESY, Hamburg and Zeuthen, Germany

E-mail: andreas.haupt@desy.de, andreas.gellrich@desy.de, yves.kemp@desy.de, kai.leffhalm@desy.de, dmitri.ozerov@desy.de, peter.wegner@desy.de

Abstract. DESY is one of the world-wide leading centers for research with particle accelerators, synchrotron light and astroparticles. DESY participates in LHC as a Tier-2 center, supports on-going analyzes of HERA data, is a leading partner for ILC, and runs the National Analysis Facility (NAF) for LHC and ILC in the framework of the Helmholtz Alliance, Physics at the Terascale. For the research with synchrotron light major new facilities are operated and built (FLASH, PETRA-III, and XFEL). DESY furthermore acts as Data-Tier1 centre for the Neutrino detector IceCube.

Established within the EGI-project DESY operates a grid infrastructure which supports a number of virtual Organizations (VO), incl. ATLAS, CMS, and LHCb. Furthermore, DESY hosts some of HEP and non-HEP VOs, such as the HERA experiments and ILC as well as photon science communities. The support of the new astroparticle physics VOs IceCube and CTA is currently set up.

As the global structure of the grid offers huge resources which are perfect for batch-like computing, DESY has set up the National Analysis Facility (NAF) which complements the grid to allow German HEP users for efficient data analysis. The grid infrastructure and the NAF use the same physics data which is distributed via the grid.

We call the conjunction of grid and NAF the DESY Grid Centre.

In the contribution to CHEP2012 we will in depth discuss the conceptional and operational aspects of our multi-VO and multi-community Grid Centre and present the system setup. We will in particular focus on the interplay of Grid and NAF and present experiences of the operations.

1. Grid Centre overview

The DESY campus is spread over two locations. The major one is located in the city of Hamburg. The second site is placed in Zeuthen, a small town near Berlin. Both locations are separated more than 300 km of distance. This spatial separation results in some operational constraints.

The DESY Grid Centre hosts a large fraction of the DESY computing resources. It consists of three blocks: The two grid sites DESY-HH and DESY-ZN on the one hand and the interactive National Analysis Facility "NAF" on the other. DESY-HH represents the grid infrastructure of the Hamburg site, DESY-ZN is located at the Zeuthen site.

The grid sites are operated individually by the local computing centre groups. The NAF on the other hand is operated as a joint effort between the two DESY computing centres in Hamburg and Zeuthen.

Published under licence by IOP Publishing Ltd

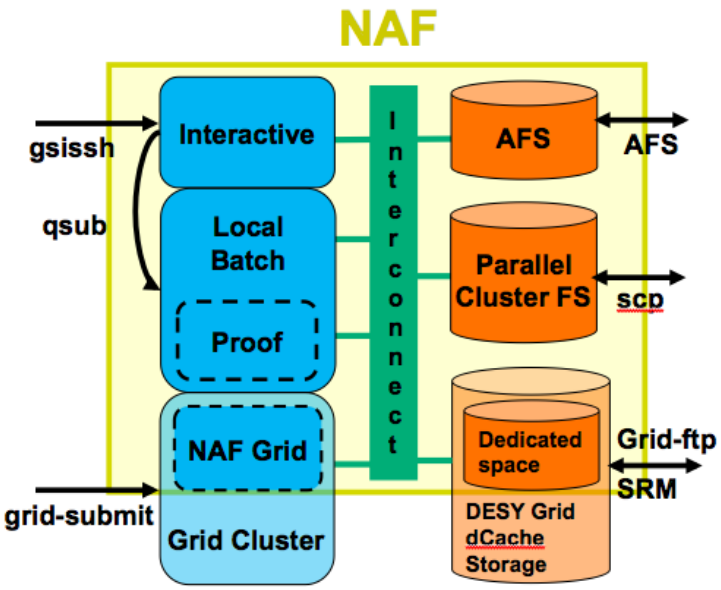

Figure 1: Schematic NAF

2. The Grid Centre in detail

2.1. The National Analysis Facility - NAF

The NAF was established within the framework of the German Helmholtz Alliance, "Physics at the Terascale"[1] which was founded to bundle the German activities in the field of high-energy collider physics.

The NAF is meant as an interactive add-on to the large-scale data and computing resources already provided inside the grid infrastructure. The focus lies on a fast and easy access to data hosted at DESY hiding the complexity of the grid. Significant computing resources, managed by a Gridengine batch system, have been installed to enable users to finish analysis tasks on a short timescale. Furthermore a workgroup server infrastructure has been established. A schematic overview is availabe in figure 1.

More details have been published already at CHEP 2010 in [1].

2.2. Storage services

The main goal of the Grid Centre is to provide a large-scale, high-performance and high-available storage service. As there is currently no technical solution available that covers all use cases, different storage systems have to be provided. They include:

- dCache
- Lustre
- Sonas (evaluation)
- AFS

Benchmarks comparing the different storage solutions with the aspect on typical LHC analysis tasks were done and described in [2].

[1] http://www.terascale.de/

2.2.1. dCache: dCache is the main storage solution for managing data import and export. It provides several data access protocols natively like e.g dCap, GridFTP, NFS4.1 and xrootd. Therefore it perfectly integrates in many experiment frameworks. For scaling and redundancy reasons, six independent dCache instances providing more than 5 Petabyte of storage capacity in total are consolidated within the Grid Centre. Each of them provides storage for different experiments:

- Atlas (Hamburg)
- CMS (Hamburg)
- "DESY" - ILC and other non-LHC VOs (Hamburg)
- Atlas (Zeuthen)
- LHCb (Zeuthen)
- Astroparticle and Theory (Zeuthen)

2.2.2. Lustre: Lustre was chosen as an additional fast scratch file system optimized for single client performance. The Hamburg site is currently evaluating Sonas as a replacement. Both storage solutions (Lustre and Sonas) are only visible (i.e. mounted) inside the interactive NAF.

2.2.3. AFS: AFS is the work horse in the interactive NAF. It hosts e.g. the home directories as well as experiment software. Direct access to the NAF AFS cell from home institutes is explicitely wanted. Users with a working AFS client on their end system can access the AFS exported data.

2.2.4. Sonas: In the Hamburg site of the NAF, Lustre is currently being replaced by IBM Sonas. The Sonas system has been designed for higher throughput, should offer higher stabiliy. It furthermore provides tools to manage quotas and to generate reports. Additional services based on the IBM Sonas cross-cluster mount technology like GridFTP access to Sonas are under investigation with IBM.

2.3. Computing service
Each of the "Grid Centre"-blocks operates an independent batch system. Whereas Torque is used on the Grid infrastructure at DESY-HH and DESY-ZN, Gridengine handles the interactive batch resources in the NAF.

Due to partly different use cases, different approaches on the computing-service setup have been realised. At DESY-HH e.g. physical CPU cores in the batch system are oversubscribed[2] in order to maximize the overall job throughput: hyperthreaded CPUs are actively in use to some extend. As RAM is the main cost factor in current procurements, only 3-4 GB per "real" physical CPU core are available in a compute node for a reasonable price. Due to this constraint an average job on oversubscribed CPUs is restricted to 2-3 GB memory usage. This still meets the current requirements in WLCG[3].

On the other hand some non-LHC communities (e.g. Astroparticle) often have a higher demand on RAM per job. CPU oversubscription would contradict this requirement. This is the reason why it is not in use at DESY-ZN. As a side effect this leads to a higher average HS06[4]-benchmark rating and also shorter average job runtimes. As a deficit this approach leads

[2] all hyperthreaded logical CPU cores are being used - not just the physical ones
[3] Worldwide LHC Computing Grid
[4] HEP-SPEC: https://hepix.caspur.it/benchmarks/doku.php

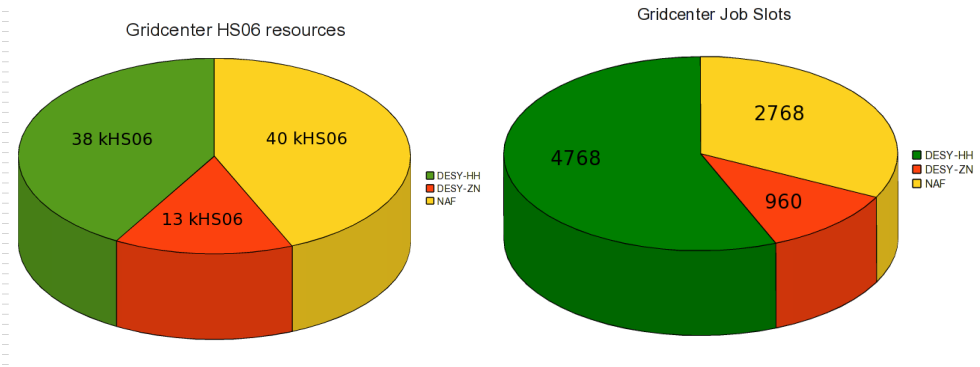

Figure 2: Comparison of the Grid Centre computing resources based on HS06 rating and job slots

to slightly higher unused compute cycles. In the NAF CPU oversubscription is just allowed for short-running jobs.

A comparison of the provided computing resources (based on job slots and HS06 rating) is shown in figure 2. In total the Grid Centre provides more than 90 kHS06 spread over more or less 8500 job slots.

2.4. Operational details / issues

The DESY Grid Centre operates many central services for VOs hosted at DESY like "zeus", "hone", "hermes", "icecube", "ilc" and "calice". Services provided include LFC[5]s, VOMS[6] servers, a redundant WMS[7] farm and top-level BDII[8]s. They are mainly located inside the DESY-HH grid infrastructure.

The DESY grid services are currently in the process of being migrated to the new EMI[9] middleware. This includes e.g. the BDII, VOMS and LFC service. Worker nodes still run with gLite-3.2 due to issues of VO frameworks with the new middleware. These are already addressed so that a migration hopefully can start soon.

A new approach of VO-software distribution is the CernVM File System (CVMFS[3]). Based on a hierarchical caching infrastructure the scaling issues with current central file servers are avoided. Even a high-availability service is built in by design. On the other hand this software is still in an early stage of development and VO administrators as well as ourselves (the grid infrastructure admins) have to get familiar with it. As an example Atlas was rather unspecific on the minimum quota of the different repositories so that they are functional. Typically there was a far too high requirement for the limited and shared resource "disk space" on a node. The technical possibility of setting a quota per VO (and not just per repository) might help here. Nevertheless CVMFS has already been deployed successfully in the NAF as well as at DESY-ZN and replaced the AFS/NFS-based software installation for some VOs.

As already shown the DESY-HH batch system hosts more then 4500 job slots. Due to limitations in the Maui scheduler (part of the EMI/gLite middleware), it does not scale with the job throughput. This led to unused job slots as Maui was not able to schedule as fast as

[5] Logical File Catalogue
[6] Virtual Organization Membership Service
[7] Workload Management System
[8] Berkeley Database Information Index
[9] European Middleware Initiative

jobs finished. In order to tackle this deficit Andreas Gellrich wrote a fast and simple scheduler replacement (see: [4]).

3. VO support

3.1. Tier0 for the HERA and ILC experiments

Although the HERA[10] accelerator was switched off some years ago, its data processing is still going on. VOs like hone (H1), zeus and hermes are actively using grid resources for simulations and data analysis. The Grid Centre acts as a Tier0 for those VOs.

Additionally the grid resources are also accessible by the ILC[11] experiments and their collaborations (ILC, Calice). DESY plays and will play the Tier0-role for them, as well.

3.2. Photon science

The DESY research portfolio is currently being dramatically enhanced by many photon science experiments. The Grid Centre is prepared for a huge increase of computing and storage demands. Present accelerator-experiments like PETRA-III or FLASH are supported as well as future ones (XFEL[12]) and their community groups (e.g. HASYLAB, CFEL).

3.3. Tier2 for Atlas, CMS and LHCb

The DESY Grid Centre acts as a Tier2 for three LHC experiments: Atlas, CMS and LHCb. Whereas the CMS-Tier2 is concentrated on the Hamburg site, the LHCb-Tier2 is located in Zeuthen. The Atlas-Tier2, on the other hand, is spread over both grid sites - having pledged a half Tier2 each.

The Grid Centre resources are not fully used by all its stakeholders all the time. Especially the Atlas and CMS VOs profit very much from the shared facility (see: figure 3). Both VOs could consume much more than the pledged DESY CPU resources - in some months even more than whole pledges for Germany, which correspond to around 10% of requested resources for ATLAS and CMS worldwide!

3.4. Icecube Tier1

The Zeuthen site operates as a Tier1 centre for the IceCube project. It hosts half of the Monte Carlo data and all of the second-level data. In addition, it acts as a backup for the Tier-0 site in Madison, USA. See also: [5]

4. Summary

The DESY Grid Centre is one of the largest grid-enabled computing facilities in the world. It supports more than 20 VOs from different user communities. All in all it provides more than 90000 HS06 in more than 8000 CPU slots. The provided storage resources currently sum up to more than 6 Petabyte. Furthermore the DESY Grid Centre offers VO-services like LFCs and VOMS servers which complete the grid-service infrastructure.

[10] Hadronen-Elektronen-Ring-Anlage
[11] International Linear Collider
[12] X-Ray Free-Electron Laser

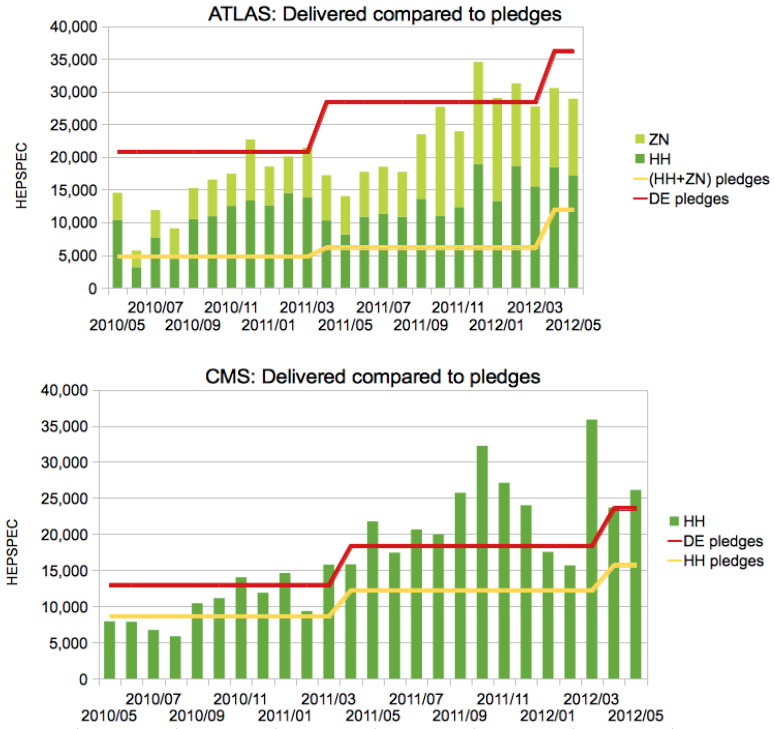

Figure 3: Comparison of pledges and used computing resources for Atlas and CMS

Reference

[1] Aplin S, Ehrenfeld W, Haupt A, Kemp Y, Langenbruch C, Leffhalm K, Lucaci-Timoce A and Stadie H 2010 *CHEP 2010 Proceedings*
[2] Fuhrmann P, Gasthuber M, Kemp Y and Ozerov D 2012 Experience with HEP analysis on mounted filesystems These proceedings (CHEP2012)
[3] de Salvo A 2012 Software installation and condition data distribution via CernVM FileSystem in ATLAS These proceedings (CHEP2012)
[4] Gellrich A 2012 Optimizing Resource Utilization in Grid Batch Systems These proceedings (CHEP2012)
[5] Barnet S 2012 The IceCube Computing Infrastructure Model These proceedings (CHEP2012)