Performance optimization for the present and next generation HEP data analysis on the Grid

as part of the Helmholtz Alliance

"Physics at the Terascale"

Proposal coordinated by the Grid Project Board

Introduction:

The overwhelming success of the LHC physics programme, culminating in the recently announced discovery of a new particle with properties as expected for a Standard Model Higgs boson, was also possible because of the excellent performance of the LHC Computing Grid. On an international level, German sites ranked among the most-used and most-efficient sites in the WLCG. The success of the Grid Project was explicitly acknowledged in the Mid-Term review of the Helmholtz Alliance. To quote from the report: "These actions have engaged an increased number of university groups directly with the WLCG activities, thereby increasing the German expertise and experience in this area. This will be vital to Germany fully exploiting the LHC."

The Grid Project of the Helmholtz Alliance "Physics at the Terascale" has significantly contributed to the Grid infrastructure in Germany for simulation, reconstruction and analysis of LHC data. Major contributions by the Alliance were made in the area of Tier-2s at the university sites Aachen, Freiburg, Göttingen, München and Wuppertal, in close collaboration with DESY and the Max-Planck-Institute for physics in Munich, operating Tier-2 centres for ATLAS, CMS and LHCb, and KIT, the German Tier-1 site, which also provides dedicated resources for national users.

The resources in Tier-2 centres at universities presently cover more than 50% of the total Tier-2 capacity available in Germany. The provisioning of personnel, infrastructure for computing hardware, for networking and the costs for electric power and cooling by the operating institutions complemented the funding for hardware through the Alliance. Significant contributions were also provided through the German Ministry of Education and Research (BMBF) for experiment-specific tasks at Tier-2 sites, which include the installation of experiment software, user support, user management and the communication with and coordination within the collaboration-wide computing groups. Although there will be no further funding of hardware by the Alliance in the years 2013 and 2014, all sites have agreed to continue operating the installed and still usable hardware. BMBF support for experiment-specific tasks also continues, although at a reduced level.

Through the Grid project of the Alliance, support was provided to all German Grid sites for Grid-based mass storage (dCache) and monitoring of site performance. Very visible contributions were further made in virtualisation of resources, and application-driven monitoring of user jobs on the Grid. These research projects of the alliance with contributions by all German WLCG sites stimulated a rich research programme on computing in Germany and provided the necessary expertise for coverage of the operational work needed to efficiently operate the Grid infrastructure and to take care of the experiment-specific services.

In the years 2013 and 2014, the interim funding of the Helmholtz Alliance will play a crucial role in ensuring Germany's position at the forefront of computing developments in High-Energy Physics and for the preparation of the computing infrastructure necessary to cope with the challenges after the restart of the LHC in the year 2015 following the upgrade to a centre-of-mass energy of 14 TeV during the years 2013 and 2014. It will be essential to keep the expertise built up by the groups and enable them to continue to contribute to the successful operation of the Grid infrastructure in Germany. Through the universities, urgently needed young academics for the future planning, implementation and operation of computing infrastructure in a rapidly changing environment can be acquired.

The requirements for technical data analysis tools for the Tier 2 centres and the NAF are growing strongly. Based on experience with data analysis the analysis concepts of the experiments will evolve and become more elaborate. The efficient management of large amounts of data and their analysis requires the deployment of further developed storage solutions. New search and access methods for data reduction, strategies for efficient utilization of fast and reliable networks, and tools for dynamic data management and use of data caching must be developed. The Grid system software and analysis tools need new technical developments, such as cloud computing in various forms whose use must be extended so that they can be used successfully.

As it will be impossible to fund the full width of computing activities required by the Alliance during the interim period, this proposal focuses on key projects with important impact on the future of distributed computing in High-Energy Physics. Specific areas of prominent importance and with strong expertise by German groups were identified for this proposal:

- Development of reliable and high performance access to LHC data through the dCache project,
- Performance monitoring of Grid jobs, sites and services,
- Virtualisation techniques and the management of job submission and workflows on the Grid infrastructure, and exploration of new technologies like cloud computing, for which concepts exist to exploit such non-WLCG resources for particle physics applications.
- Improvement of networking connections in Germany and the international connectivity.
- General support for site operations, training and schools.

The work packages and the partners as well as the resources needed are described in the following sections.

Proposed Work Packages:

WP1: Reliable and high-performance access to LHC experiment and user analysis data at the German Grid sites (Aachen, DESY, Wuppertal)
Worldwide the majority of the WLCG Grid production sites very successfully use the dCache[1] software package to operate data storage with petabytes of data. In Germany the Tier-1 and all of the CMS and ATLAS Tier-2 centres are using dCache.

The dCache project is mainly coordinated by DESY. During the last five years the HGF Terascale Alliance supported the project with personnel. Therefore expert knowledge exists directly at the German WLCG production Tier-1 and Tier-2 sites to run dCache and to provide the experiments' production teams and German users with a very reliable access to the data with excellent performance. It is essential to further support the dCache project with personnel in order to be able for all German Tiers to stay in close contact to the project and to quickly fix storage system problems at sites, which otherwise would severely block access to the LHC data.

WLCG Grid computing is not a static project. Especially storage, one of the major components of the software package, has to be continuously adapted to the needs of the experiments in changing computing models and to the demands of the users analysing the data.

We apply for three 50% FTE positions for this Work Package for the years 2012 and 2013. Funding from the institutes or third party funding will provide the second half of each position. The partners asking for manpower funding and their planned contributions are:

DESY
With the most recent changes in the data management model of WLCG, moving away from the initially implemented MONARC model towards a „Mesh network topology“ and with the more strict separation of tape and disk based storage, the dCache project is facing new challenges in competing with storage solutions provided by other countries in terms of supported protocols, tape – disk interaction and monitoring. In order to cope with those new requirements of the LHC experiments[2], significant development work has been started. However, with the end of the EMI[3] (European Middleware Initiate) funding period next year, following up on those activities will become hard for the dCache team at DESY, where most of that work is done. The work proposed is to make dCache part of the planned storage federation and on integrating the corresponding monitoring mechanism.

[1] Large Storage Systems, present and future, Andreas Peters, Chep 2012, New York
[2] New computing Models, Ian Fisk, Chep 2012, New York,
[3] EMI, The European Middleware Initiative, http://www.eu-emi.eu/

RWTH Aachen

Over the last five years RWTH Aachen was very actively integrated in the dCache development and support project. Aachen is proposing to continue giving dCache support to all German HEP Grid sites, especially by providing hands-on tutorials for dCache administrators during the annual GridKa School and special dCache workshops where new features of dCache are presented. One part of the contribution would be to assist the German dCache sites solving problems with their dCache setup. In addition, new dCache versions and features can be fully tested only on a production system with full load. RWTH Aachen's Tier-2 offers the possibility to install dCache development versions on a large production system to debug problems and to give important fast feedback to the dCache developer team.

University of Wuppertal

dCache is a software layer, which acts on a file system. To get the best performance for the access to the data an evaluation of the optimal file system is crucial. While most sites use standard Linux file systems to operate dCache, we aim to evaluate the advantages of using cluster file systems as underlying file systems to gain performance and reliability and robustness against dCache pool node failures. A setup of commonly used cluster file systems (at least Lustre, PanFS[4], FHGFS[5]) will be tested and the performance will be measured in standalone mode and with dCache, tuning different setup and operations parameters. In addition to the benefit for the Grid sites the experience gained about optimal file systems for HEP data analysis will also provide important information for the setup of smaller institutes' data analysis clusters.

All German Tier-2 sites for ATLAS and CMS (Aachen, Freiburg, Göttingen, DESY Hamburg, DESY Zeuthen, München, Wuppertal) and the Tier-1 at KIT will benefit from this project and will cooperate without applying for dedicated manpower.

The positions requested would allow continuing to provide excellent access to the data relevant for the LHC experiments and all German analysis users who store the majority of their analysis data at our German Grid sites.

Total funding request: 3 x ½ positions for Aachen, DESY and Wuppertal

WP2: **Performance Optimization, Site and Meta-Monitoring** (Göttingen)

Performance monitoring and optimization of both Grid sites and of individual services offered across multiple sites are essential in the complex, distributed computing environment of the world-wide LHC computing Grid. To facilitate this challenging task and in particular to allow non-experts to identify actual problems rapidly, the meta-monitoring package HappyFace[6] has been developed by partners from Karlsruhe, Göttingen, Hamburg and Aachen and is presently in use at the German Tier-1 and all German Tier-2s in CMS, by CMS central operations for direct batch system monitoring at all CMS Tier1s and is also deployed at some ATLAS Tier-2s in Germany.

The KIT group has driven the HappyFace core development in the past with partial funding by the Helmholtz Alliance. Personnel from all sites performed dedicated module development for special monitoring tasks. Thus, a rich set of ready-to-use modules exists to include in the HappyFace system the monitoring of e.g. the usual batch systems, usage and access patterns of dCache storage and data transfers between sites in the HappyFace system. The core software and many of the most important modules are presently re-implemented in HappyFace version 3 to ease packaging of core components and modules, release management and distribution.

HappyFace has the potential to be extended from a passive meta-monitoring tool to an expert system that analyses the history and correlation of acquired monitoring data allowing to precisely identify and classify problems, and in addition to automatically trigger predefined actions. This will reduce site downtimes, increase site efficiencies, and at the same time reduce the amount of personnel needed to operate sites, systems or services. The classification will be implemented using Fuzzy Sets and the decision-making system will be based on

[4] Panasas: high performance parallel storage

[5] Fraunhofer Parallel Cluster File System

[6] https://ekptrac.physik.uni-karlsruhe.de/trac/HappyFace/, and V. Büge, V. Mauch, A. Burgmeier et al., CHEP 2009, Taiwan

neural networks. The system will be able to learn decisions from site administrators. Once, it faces a similar problem it can either automatically solve it or send a detailed notification to the responsible expert.

As the amount and complexity of services and monitoring sources in WLCG permanently grow, small as well as large sites would benefit substantially from this kind of automation.

Total funding request: ½ position for Göttingen

WP3: **Job and workflow Management in distributed environments** (KIT-Süd, München)
The groups from KIT (CMS) and LMU München (ATLAS) have gathered profound experience in the management of jobs and workflows in complex, distributed computing environments and developed or significantly contributed to widely used products, like GANGA for distributed analysis[7], a work-load management system based on glideins[8] via the Condor batch system, the tool HammerCloud[9] for performance tests and problem finding, and virtualisation techniques to access computing resources outside the WLCG, as provided through local batch systems at university clusters (ViBatch) or cloud resources (the meta-scheduler ROCED), see[10] .

Based on this experience, the groups in München and Karlsruhe plan to continue the development of tools to facilitate physics analysis in the rapidly changing distributed environment, largely driven by cloud computing and the availability of resources accessible through some variant of a cloud API.

KIT Karlsruhe
The group has made very significant contributions to the new workload management system deployed by CMS (glidein-WMS). The system ensures interoperability between different flavours of Grid middleware and other resources and provides a high level of scalability. The different resources are presented to the users via the Condor batch system, which thus lets appear a heterogeneous computing environment as a single, homogeneous pool of resources. This mechanism provides an effective shielding of the user from the underlying middleware.

Within this work package a glidein-WMS for usage by the German CMS groups will be implemented on dedicated hardware, which is already available at KIT. Extensions will be implemented to comfortably include standard WLCG sites as well as the German resources for physics analysis, the NAF at DESY and the national resources at GridKa; access to private clouds, as already available as test systems at KIT, and eventually commercial cloud resources will also be included. Usage of private and commercial clouds as either remote nodes of a batch system or as part of a remote batch server have already been demonstrated in two diploma theses at KIT, and a prototypes system accessing a private cloud at KIT is running already. The introduction of glidein-WMS will lead to homogeneous system by presenting all available resources through a single interface.

If successful and justified by the user acceptance of the system, a national workload management system can be set up based on the experience gained with the prototype at KIT. Experienced personnel at KIT is available to carry out the project, but help in form of ½ co-funded positions is needed to guarantee longer-term operation and support for non-KIT users.

LMU München
The LMU group played a central role in the development of the Ganga distributed analysis submission system and the HammerCloud job testing facility for ATLAS. The glidein-WMS is an attractive alternative for job submission also for the ATLAS experiment; in particular it offers more advanced features for user authentication when compared to the currently used submission based on condor. A central goal of the project will be the integration of glidein-WMS into the ATLAS frameworks Panda and Ganga.

[7] J. Mosciki et al, Ganga: a tool for computational-task management and easy access to Grid resources, Computer Physics Communications, Volume 180, Issue 11, (2009).
[8] I. Sfiligoi, M: Zvada et al, CHEP 2012, New York, USA
[9] D. van der Ster et al, Experience in Grid Site Testing for ATLAS, CMS and LHCb with HammerCloud, Proceedings CHEP 2012, New York, USA
[10] T. Hauth, O. Oberst, A. Scheurer et al., CHEP 2010, Taiwan, and CHEP 2012, New York, USA)

Further development for Ganga and HammerCloud is also required to facilitate WAN access to data files. Several projects have been started to develop services, which flexibly redirect data access to remote sites, e.g. based on the Xrootd redirector[11]. Job Management systems such as Ganga and the HammerCloud testing facility need to be adapted in order to integrate these services and fully exploit the features.

Funding request: 2 x ½ positions for Karlsruhe and München

WP4: Wide-area network in Germany and international connectivity (DESY, KIT-Nord)
Connectivity is obviously an essential part of LHC computing. With LHCOPN and LHCone two network infrastructures for the international Tier-1 and the Tier-2 level where designed and successfully deployed to the benefit of LHS data analysis in operation. GridKA covers the financing of the Tier-1 connection to LHCOPN whereas the connection of some of the Tier-2's in Germany (Aachen, Wuppertal, Hamburg) and GridKa is currently partially financed by the Alliance. In particular the 10 Gbit/s interface from the national part of the LHCone into the pan European GEANT Network is presently part of the Alliance funding scheme. RWTH Aachen, KIT and DESY are now largely contributing from base funds to the LHCone. Future funding of the German LHCone contribution is currently open and different funding resources are under investigation. The operational effort is considered as to be part of the standard networking operations of the centres, therefore no extra personnel is requested.

Funding request: none, work is part of the standard networking operations of the centres.

WP5: General support for site operations, training and schools (KIT-Nord, all sites)
The cooperation between the participating sites is a very important building block of the computing infrastructure to LHC data analysis. It will be essential to continue the support, organizing schools and workshops in the context of the computing project of the "Physics at the Terascale". Due to limited finances the effort will be provided by the Helmholtz centres, counting on the contributions of all participating partner sites, without the request for additional funding. The German dCache Support Team consisting of DESY, KIT and Aachen members provides support to German dCache users. Further more, the knowledge gathered within this project will be spread beyond the "Physics at the Terascale" scope in schools and workshops like the annually international GridKa School at KIT.

Funding request: none.

Resource Planning:

Funding of research positions is requested at the level of postdocs positions for the duration of the project as listed in the in the work package description above.

Work package	**Centre/Partner**	**Personnel financed by Alliance**	**Personnel financed through institute or third party funding**
WP1: Grid-enabled storage systems	RWTH Aachen	½	½
WP1: Grid-enabled storage systems	DESY	½	½
WP1: Grid-enabled storage systems	Wuppertal	½	½
WP2: Performance Optimization + Monitoring	Göttingen	½	½
WP3: Job and Workflow Management	KIT-Süd	½	½
WP3: Job and Workflow Management	LMU München	½	½
WP4: Wide-area network in Germany and international connectivity	DESY, KIT-Nord	0	0
WP5: General support for site operations, training and schools	KIT-Nord all sites	0	0
Total requested positions		**6 x ½**	**6 x ½**

[11] B. Bockelman et al, Using Xrootd to Federate Regional Storage, CHEP 2012, New York, USA

List of participating Institutes

Participating Helmholtz Centres	**Location**	**Group Leader**
Deutsches Elektronen Synchrotron	Hamburg & Zeuthen	V.Gülzow, P.Fuhrmann
Karlsruhe Institute of Technology	Karlsruhe	G.Quast
Participating Universities		
Rheinisch-Westfälische Technische Hochschule Aachen	Aachen	Th.Kreß
Georg-Augstus-Universität Göttingen	Göttingen	A.Quadt
Ludwig-Maximilian-Universität München	München	G.Duckeck
Bergische Universität Wuppertal	Wuppertal	T.Harenberg