Grid Computing at GSI for ALICE and FAIR - present and future

Kilian Schwarz, Florian Uhlig, Radoslaw Karabowicz, Almudena Montiel-Gonzalez, Mykhaylo Zynovyev, Carsten Preuss

GSI Helmholtzzentrum für Schwerionenforschung GmbH, Planckstraße 1, 64291 Darmstadt, Germany

E-mail: k.schwarz@gsi.de

Abstract. The future FAIR experiments CBM and PANDA have computing requirements that fall in a category that could currently not be satisfied by one single computing centre. One needs a larger, distributed computing infrastructure to cope with the amount of data to be simulated and analysed.

Since 2002, GSI operates a tier2 center for ALICE@CERN. The central component of the GSI computing facility and hence the core of the ALICE tier2 centre is a LSF/SGE batch farm, currently split into three subclusters with a total of 15000 CPU cores shared by the participating experiments, and accessible both locally and soon also completely via Grid. In terms of data storage, a 5.5 PB Lustre file system, directly accessible from all worker nodes is maintained, as well as a 300 TB xrootd-based Grid storage element.

Based on this existing expertise, and utilising ALICE's middleware 'AliEn', the Grid infrastructure for PANDA and CBM is being built. Besides a tier0 centre at GSI, the computing Grids of the two FAIR collaborations encompass now more than 17 sites in 11 countries and are constantly expanding.

The operation of the distributed FAIR computing infrastructure benefits significantly from the experience gained with the ALICE tier2 centre. A close collaboration between ALICE Offline and FAIR provides mutual advantages. The employment of a common Grid middleware as well as compatible simulation and analysis software frameworks ensure significant synergy effects.

1. GSI in the context of ALICE and FAIR

GSI (GSI Helmholtzzentrum für Schwerionenforschung GmbH) [1] operates a large and in many aspects worldwide unique accelerator facility and employs more than one thousand people. Researchers from Europe and from around the world conduct experiments here extending from nuclear and atomic physics to plasma and materials research, and encompassing biophysics and cancer therapy.

GSI maintains several local experiments, e.g. HADES [2], but also participates in the ALICE [3] experiment at CERN. GSI and the surrounding universities are responsible for building the detector components TRD (Transition Radiation Detector) and TPC (Time Projection Chamber) of the ALICE detector. At the same time the GSI related groups are developing corresponding software and are doing calibration and data analysis, mainly related to the local detector components. Moreover GSI operates a tier2 centre within the ALICE computing Grid.

Centered on GSI, in the years to come an international structure named FAIR (Facility for Antiprotons and Ion Research) (see section 3.1) will evolve. The international FAIR GmbH (limited liability company law) was founded on October 4, 2010 and the first beam is expected by 2018.

Up to now the investment in computing at GSI was dominated by the needs of the ALICE experiment but is gradually shifting towards FAIR requirements.

2. GSI Computing Today

2.1. The GSI computing infrastructure

At the core of the GSI computing infrastructure is the large compute cluster (batch farm) which is currently distributed over three sub clusters which are located at different places in the GSI campus. The total size is about 15000 compute cores. The old farm consists of 340 nodes or 2700 cores and is the only farm which still has NFS based shared directories which are equally visibile from the submit hosts and the worker nodes. Moreover it is the only part still accessible via Platform LSF. Meanwhile the transition from the commercial LSF scheduler to the open source scheduler (Sun)GridEngine [4] has been accomplished. Other scheduling systems have also been tested beforehand as e.g. OpenPBS/Torque which was not chosen due to lack of stability and scalability. After Sun has been bought by Oracle in 2011 the future of (Sun)GridEngine became complex, though. Unlike before currently there are several development branches and support communities, some of them commercial. This implies the questions which (Sun)GridEngine software stack to install and which development branch is going to be supported in future on what basis. Also some technical problems still exist, e.g. issues with enforcing resident memory limit. The 2 new clusters, an intermediate test cluster and the so called MiniCube cluster (in total 12000 cores) which serve as a test case for the upcoming Green Cube (section 3.2) computing centre at GSI have been set up using exclusively GridEngine and several more new concepts. Concerning scheduler configuration the system has been moved from queue specific scheduling to resource specific scheduling. Resource requests are moreover being used as resource limits, i.e. jobs exceeding the given limits are being killed. This way the cluster utilisation increased by 5-10%. Due to scalability reasons NFS mounts and NFS based shared directories are not being used anymore. Experiment software packages used for data processing and analysis are being installed and distributed on the farm using the network file system cvmfs [5] which is based on http and developed at CERN. For Infrastructure Management it has been decided to use the Ruby based Chef [6] system. The mass data are to be stored on Lustre [7] as was the case before. In total Lustre capacity of 5.5 PB is being provided distributed over two clusters, one being visible from the old GSI infrastructure and one being visible from the MiniCube cluster. The network is based exclusively on InfiniBand with clearly defined boundaries to the remaining Ethernet based GSI infrastructure.

For its energy and cost saving concept the GSI MiniCube cluster received the German Computing Centre Price 2012. The basic idea is to use a minimum of energy and space. Therefore the 96 compute racks were arranged like in a high rack warehouse. The idea of the cooling concept is that the warm air at the backside of the closed racks in which the computers are located is being cooled down directly with water by using heat exchangers. The actual cooling then happens via heat of evaporation. Using this technique a PUE (power usage effectiveness) of less than 1.07 could be achieved.

A schematic overview of the current GSI computing infrastructure can be seen in figure 1.

2.2. The ALICE T2/T3 centre

Together with the GSI/EMMI ALICE group GSI-IT is responsible for the operation of the ALICE tier2/3 computing centre at GSI. The ALICE tier2/3 centre provides computing infrastructure for ALICE Grid and an analysis platform to all ALICE members in Germany.

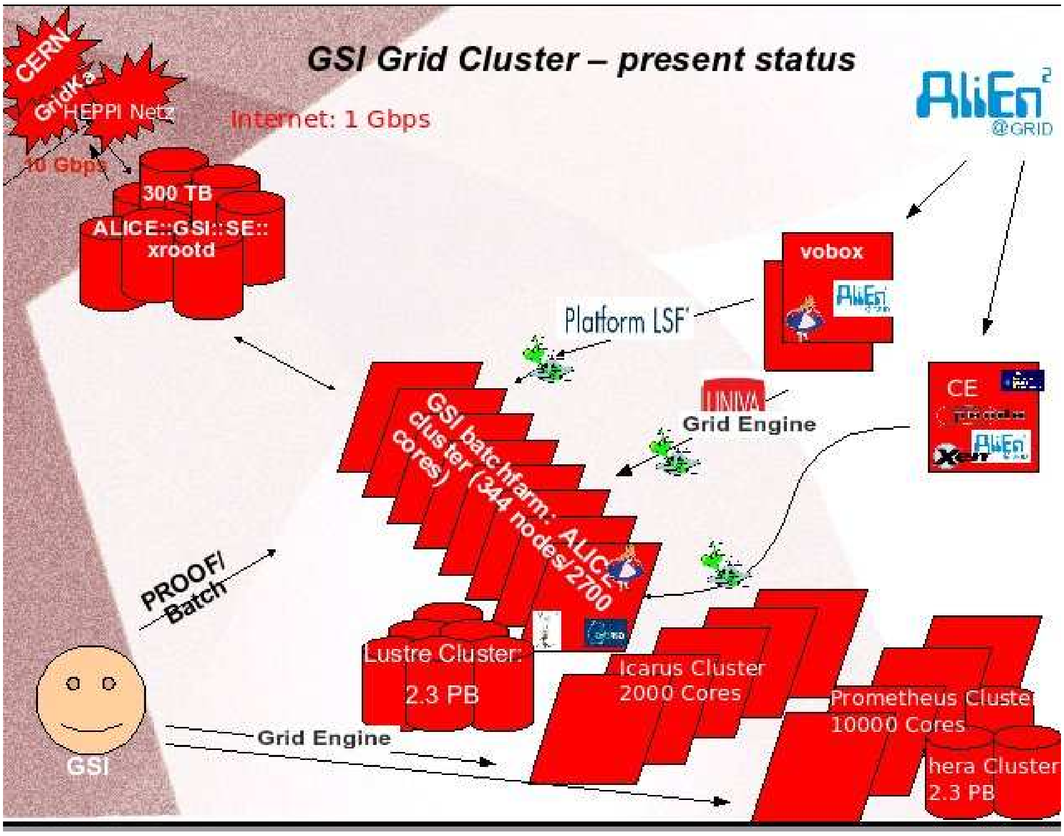

Figure 1. A schematic overview of the GSI computing infrastructure.

One of the consequences is that new data-sets are being transferred to GSI almost continuously. In order to achieve the necessary data rates GSI has been integrated with a 10 Gb link into the the HEPPI net structure of the LHC tier2 centres in Germany. The data are being processed on the local batch farm via daily running analysis trains. A train is assembled from a list of modules that are sequentially executed and is an efficient way to process a large set of data.

The available disk space at the GSI tier2/3 is distributed among an xrootd [8] cluster of 300 TB capacity for usage from the Grid and Lustre (see above) for local storage. In order to achieve synergy effects a Grid storage element consisting of an xrootd daemon running on top of Lustre has been set up. As soon as sufficient disk space is available the new setup will go into production mode.

The whole infrastructure, the local cluster as well as the whole ALICE grid infrastructure in which the ALICE tier2 centre at GSI is embedded, is monitored in detail by using MonALISA [9]. The MonALISA system is an ensemble of agent-based subsystems registered as dynamic services, that are able to collaborate in performing a wide range of information gathering and processing tasks. MonALISA provides agents to supervise applications, restart or reconfigure them and to notify other services when certain conditions are detected. Moreover it provides a very user-friendly and highly customizable Graphical User Interfaces to visualize complex information.

2.2.1. ALICE Grid in Germany

The ALICE Grid resources in Germany are provided by the tier1 centre at GridKa in Karlsruhe and the tier2 centre at GSI. Moreover the HHLR computer cluster at Goethe University in Frankfurt has been integrated in 2011. Throughout the year GSI and the other centres participate in centrally managed ALICE Grid productions and data analysis activities. During the last year the overall job share of the German ALICE Grid centres have been 17% of all ALICE jobs running worldwide. The job distribution among GridKa, GSI (SGE and LSF), and Frankfurt as well as the total job share can be seen in the figures 2 and 3.

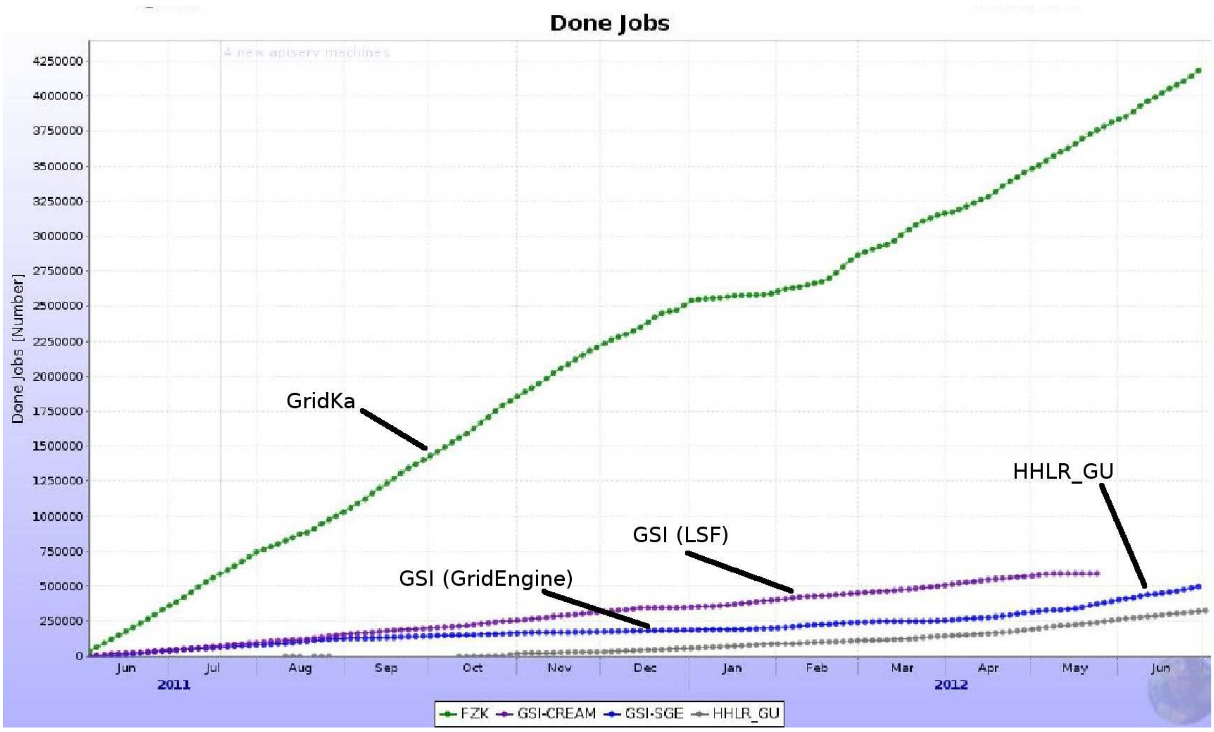

Figure 2. Here the ALICE Grid jobs successfully executed within the last year at the German Grid sites GridKa, GSI (LSF and Grid Engine Clusters), and Frankfurt (HHLR-GU) can be seen. In total alone at GridKa more than 4000000 jobs have been executed.

2.3. Preparation for FAIR

A distributed computing infrastructure for FAIR is being set up. FAIRGrid is currently implemented as two seperate entities: PandaGrid and CBMGrid.

2.3.1. PandaGrid

PandaGrid is the agreed on computing infrastructure of PANDA. The PANDA (anti-**P**roton **AN**nihilation at **DA**rmstadt) experiment [10] started as a part of FAIR (an International **F**acility for **A**ntiproton and **I**on **R**esearch) project in 2002. Currently the collaboration switches from development into building phase. The official simulation and reconstruction software is PandaRoot, a branch of the FairRoot [11] project, which dates back to 2003. Based on the experience of the LHC experiments, the general computing trends and the future requirements of the PANDA experiment, the PandaGrid project also started in 2003.

PandaGrid uses the AliEn [12] middleware, developed by the ALICE experiment. The data analysis and simulation framework, PandaRoot, is distributed via the built-in package management mechanism of AliEn. The grid monitoring and data supervision are done via MonALISA [13]. PandaGrid is constantly evolving, in 2012 the biggest site so far was added - Mainz computing cluster with 400 CPUs dedicated to the usage within PandaGrid.

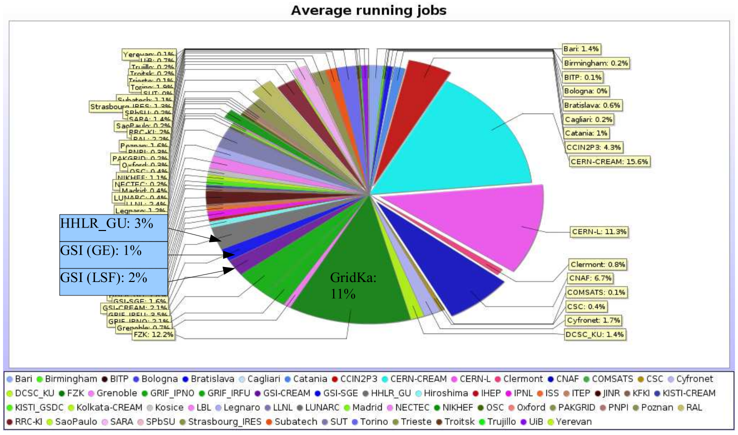

Figure 3. Job distribution among ALICE Grid sites. Within the last year 17% of all Grid jobs were running at the German Grid sites.

Additionally, the central services and main databases were transferred from Glasgow University to GSI in March 2012. This was necessary because Glasgow University left the PANDA collaboration.

Currently 17 sites from 13 institutes in 10 countries are part of the PandaGrid (see figure 4). Among them are also EGEE/EGI sites integrated via VOBox tools. The total number of CPUs available to the user is about 1000, and the storage capacity of the current PandaGrid configuration is about 100 TB.

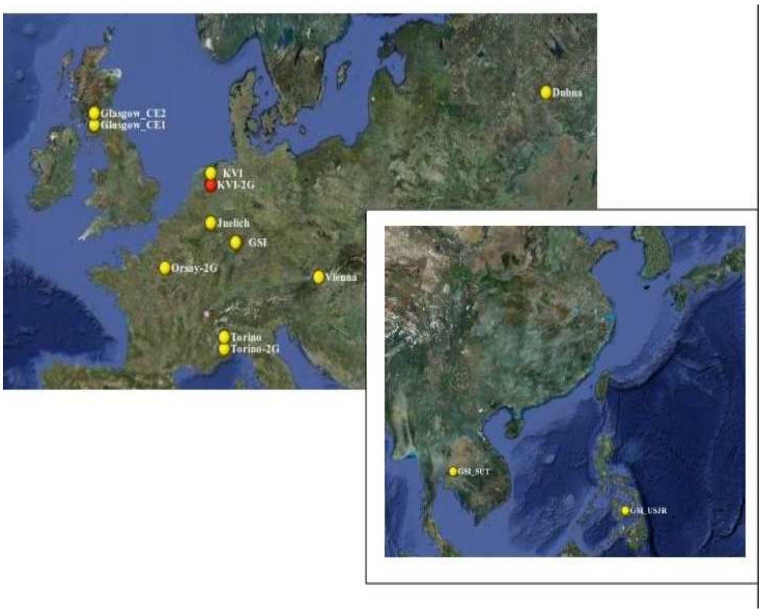

Figure 4. PandaGrid sites distributed in Europe and Asia.

One of the main advantages of AliEn is enabling compatibility between the server/worker nodes and the batch farm system regardless of the platform. That fits also perfectly to the free choice of the platform one has when using the PANDA analysis software - PandaRoot. This increases the number of potential sites and reduces the work of system administrators.

Moreover the process of the software installation is fully automatized. The process is tested every week with the newest version of PandaRoot and the results are published on the dashboard webpage [14]. To assert maximum flexibility for the users, several different versions of the PANDA software are available and the user can choose what is needed. In case the analysis is not standard there are means to upload specific analysis code to all of the sites concurrently.

2.3.2. CBMGrid

CBMGrid is up and running, but not yet in production mode. The technical implementation is identical to the one of PandaGrid (sec. 2.3.1). Also CBMGrid uses the AliEn middleware and instead of PandaRoot CBMRoot is being deployed. CBMGrid has been the first test case for the new AliEn database interface (sec. 2.3.4) First small data productions have been running successfully at GSI using the CBM Grid infrastructure. The largest CBM Grid site is Dubna, supported by the JINR-BMBF grant.

2.3.3. Resources on demand

One of the ongoing activities of the E-Science Group at GSI is development of techniques for efficient utilization of virtual resources at the IaaS (Infrastructure as a Service) clouds. Virtualization and cloud computing technologies present an opportunity to avoid software incompatibility issues between various scientific computing software and the single provided and supported platform at the GSI computing farm, currently Debian 6. For example, the FLUKA [15] software, which among other things is used for simulation of radiation at the FAIR accelerator facilities, is supported only for the Scientific Linux OS.

Besides, to meet peak demands for computing at GSI, it may be necessary to offload some of the computing tasks to public or community clouds. Tasks that are CPU bound and have negligible I/O requirements are most suitable for this. Thus, it becomes important to deploy and operate an infrastructure to compute these tasks on virtual machines in a quick and scalable way. The "infrastructure as code" (IaC) concept is adopted from the commercial world of Web operations to achieve this. Under this concept, all matters of deployment and maintenance of an infrastructure on virtual machines are implemented in code.

A property of an IaaS cloud which allows the freedom to install the complete software stack makes it possible to roll out a computing infrastructure by just executing code. Little to no manual interaction or code modification are needed, if all infrastructure elements have a remotely accessible API. Thus, the codebase, tailored to specific scientific software, may be used on any IaaS cloud in the same way. This method is used to deploy virtual clusters on clouds.

A virtual cluster is automatically assembled with the help of a configuration management system. First, a virtual machine (VM) with an installed CMS (Configuration Management Server) is provisioned. General configuration and application data are uploaded to it. Second, the rest of VMs are deployed with installed configuration management clients and assigned roles. Each VM is then pulling its configuration data automatically from the server and gets provisioned according to its role. When a software instance needs to be aware of the changing infrastructure and network context, it queries the necessary information from the CMS server which maintains an updated index of all clients and the system state. If a job scheduling system is installed and configured automatically that way, the user is then left to submit the jobs to be executed.

The methodology and codebase were developed and tested at a private cloud prototype in GSI. Subsequently, this IaC approach has since been used on the Amazon EC2 cloud and the Frankfurt Cloud [16], a community cloud for the public research sector in the Frankfurt am Main area. Among the scientific computing tasks executed on virtual clusters are the ALICE grid jobs, the nuclear structure calculations with energy density functional methods, and the aforementioned radiation protection simulation for the FAIR construction.

In another successful demonstration of the viability and efficiency of the described way to deploy virtual infrastructures, a 1000-node PROOF cluster had been deployed on a private cloud prototype at the LOEWE-CSC supercomputer [17] of the Goethe University Frankfurt. The setup of the PROOF cluster had been performed with the help of PoD [18], which used its SSH-plugin to connect the worker nodes. It allowed for the first time to benchmark the PROOF workload distribution mechanism on such a large scale installation.

2.3.4. Grid development

The AliEn middleware has been adapted for its use in the future experiments PANDA and CBM, both within the FAIR project. This software has been deployed already in 2003 for PANDA. Since then, continuous feedback has been provided, making it possible to improve the tool. Furthermore, software development has been performed at GSI. One of the lines of development has been to implement a way to make the database interface as generic as possible. This way any Relational Database Management System (RDBMS) can be enabled with fairly little changes. AliEn uses the Database Interface (DBI) Perl module for the communication with the database. Specific connections to the database come with Database Driver (DBD) modules. Originally, AliEn could be connected using exclusively the MySQL driver. As it can be seen in figure 5, a new design has been implemented. The module Database is the actual interface to AliEn. It implements the common functionality of any RDBMS, only including standard SQL statements. Due to performance and functionality reasons, it is still necessary to use particular features for each RDBMS, and this implies using specific SQL statements, which are now located in the sub-modules instead.

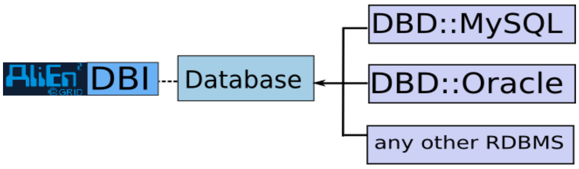

Figure 5. New AliEn database interface.

The Oracle backend for AliEn has been already developed as a first approach because it is one of the services provided at GSI. This result was already included in the AliEn v2.19 release where users could find the Oracle client bundled with the middleware.

Another achievement is the interface for the SLURM [19] batch-queuing system. AliEn Computing Elements act as interface to the local scheduler. Even though some sites are using this scheduling system, it has been never used before at ALICE Grid. With this development, an interface to this system has been enabled. Specifically, the site HHLR_GU (Hessisches Hochleistungsrechenzentrum der Goethe-Universitaet) located in Frankfurt, has been used as testing system and currently is running about 1200 jobs in production.

3. GSI Computing 2018

3.1. The FAIR project

The Facility for Antiproton and Ion Research in Europe (FAIR) [20] will generate antiproton and ion beams of a previously unparalleled intensity and quality. In the final design FAIR consists of eight ring colliders with up to 1,100 meters in circumference, two linear accelerators and about 3.5 kilometer beam control tubes. The existing GSI accelerators serve as an injector. In October 2010 FAIR was founded by means of an international agreement. About 3000 scientists and engineers from more than 40 countries are already involved in the planning and development of the facility and its experiments.

FAIR will support a wide variety of science cases: extreme states of matter using heavy ions (CBM), nuclear structure- and astrophysics (NUSTAR), hadron physics with antiprotons

(PANDA), atomic and plasma physics as well as biological and material sciences(APPA).

The high intensities at FAIR constitute various challenges: high intensities require very efficient accelerators, remote handling in activated areas, novel methods of cooling for the detectors, progress in collaborative computing and synergies between the various RI concerning the interaction with industry.

The first beam is expected to be in 2018. This constitutes a crucial milestone in computing. A schematic overview of the FAIR setup can be seen in figure 6.

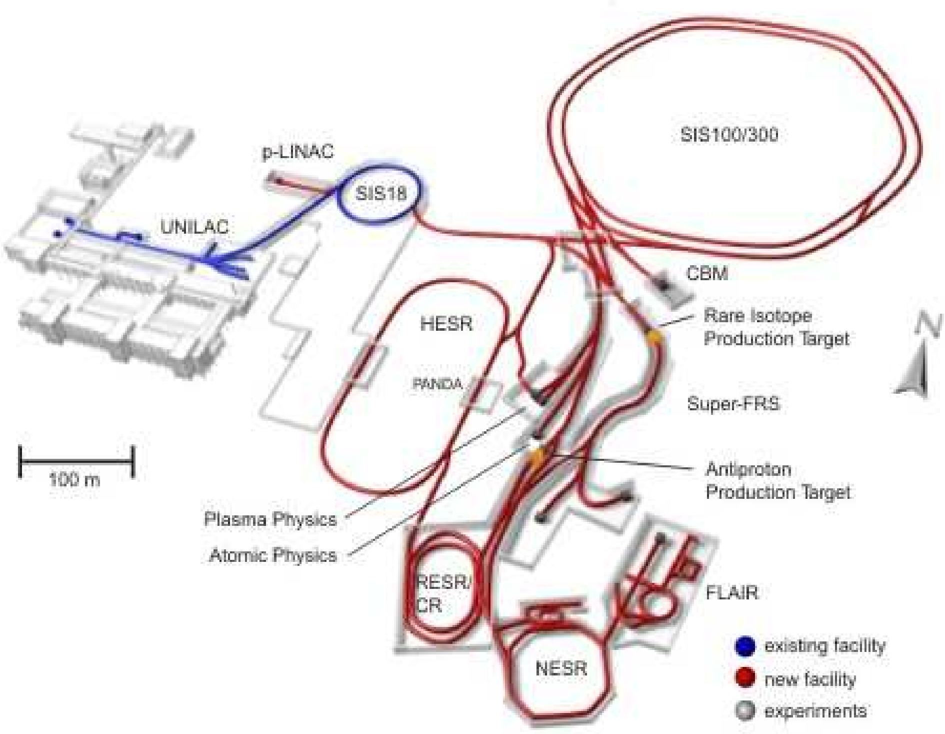

Figure 6. A schematic overview of the FAIR setup.

3.2. Characteristics of FAIR computing

For two of the research topics of FAIR (CBM and PANDA) the workflow is comparable to the one of other High Energy Physics experiments, e.g. the LHC experiments at CERN, but with certain extensions.

Triggerless data handling techniques: The current paradigm for data analysis in high-energy physics relies on trigger systems. A trigger is a system that uses simple criteria, usually derived from a subset of the detector information, to rapidly decide which events in a particle detector to transport to further data processing and archive stages. Trigger systems were necessary due to real-world limitations in data transport and processing bandwidth. At FAIR a novel triggerless detector read-out will be implemented, without conventional first-level hardware triggers, relying exclusively on event filters. This is a new approach which allows addressing physics signatures which require complex algorithms, like a full-track reconstruction or information from many detector subsystems. This approach is much more flexible and adaptable to yet unforeseeable needs because the full detector information is available even in the first decision stage, resulting in a considerably higher discovery potential. Complex filtering and flexibility are the key enabling factors for the experiments at FAIR. So at FAIR the classical separation between data acquisition, trigger, and off-line processing is merging into a single, hierarchical data processing system, handling an initial data stream exceeding 1TB/sec. The first layer of the system constitutes the first level event selector (FLES). The FLES implements a combination of specialized processing elements such as GPUs, CELL or FPGAs in combination with COTS computers, connected by an efficient high-speed network. After the FLES the data stream fans out into an archival system and into the next distributed processing layers, which can be off-site.

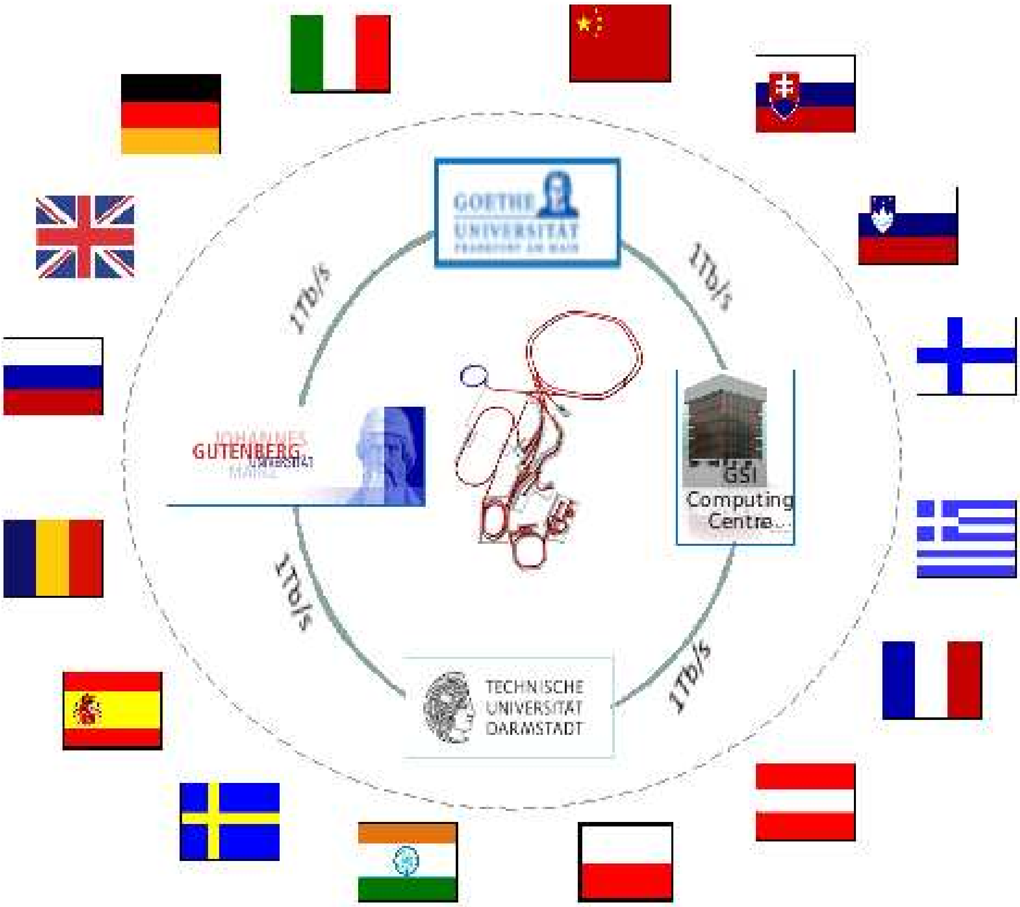
GOETHE UNIVERSITÄT FRANKFURT AM MAIN
1Tb/s
1Tb/s
JOHANNES GUTENBERG UNIVERSITÄT MAINZ
GSI Computing Centre...
1Tb/s
1Tb/s
TECHNISCHE UNIVERSITÄT DARMSTADT

Figure 7. Distributed T0/T1 centre enbedded in Grid/Cloud.

Data centres and distributed computing: Currently large-scale research infrastructures like e.g. LHC at CERN rely on on-site tier-0 data centres. Normally the tier-0 data centre performs the first-level processing after the last trigger step. Subsequently a considerably reduced amount of data is analyzed off-site in downstream data centres, operating on so-called event summary data (ESD) sets, precomputed in the tier-0 centre. In contrast to that FAIR will make use of a novel modular data processing paradigm using multi-site load balancing data centres. Several sites will be connected with a high-speed Metropolitan Area Network via fibre link allowing the off-loading of processing between the sites. That combined tier-0/1 system will be integrated in an international Grid/Cloud infrastructure (see figure 7). In the centre of the FAIR computing infrastructure will be the new computing centre, the so called "Green Cube". In a cube shaped building of a side length of 25 to 30 m computers in about 800 closed racks will be combined to become a super computer. The systems will be a mix of high performance servers, augmented with GPUs. The cooling concept is the same as described in section 2.1. This way the cooling costs will be only 8% of the needed power consumption for the computers.

4. Summary and Outview

GSI IT is engaged in LHC computing since 2001. As part of the Worldwide LHC Computing Grid an ALICE tier2 centre has been set up and is being operated at GSI. GSI is developing software and operating procedures for LHC and is actively taking part in LHC boards and committees, e.g. the Technical Advisory Boad and the Overview Board of the corresponding tier1 centre at GridKa [21]. All these combined experiences in LHC computing are being used for the setup of the tier0/1 centre at GSI and surrounding universities for the upcoming FAIR project. When comparing data rates and duty cycles FAIR computing will be in the same order of magnitude as LHC computing in terms of computing power and storage requirements. FAIRRoot started as a new implementation following the design of AliRoot and is now heavily used within the FAIR community. Many ideas from FAIRRoot go back to ALICE and CERN. FAIRGrid, as implemented for the PANDA experiment within the PandaGrid project, is in production since 2003 and now heavily used on many sites over the Globe. PandaGrid can be considered as one of the first production ready components of the PANDA experiment. The work in the CBM experiment on CBMGrid started recently but it is fast evolving while profiting from the experiences gained in ALICE and PANDA. By using Grid and Cloud technology resources can be added to the FAIR computing infrastructure from various sources.

5. References

[1] http://www.gsi.de
[2] http://www-hades.gsi.de/
[3] http://aliceinfo.cern.ch/Public/Welcome.html
[4] http://gridengine.org
[5] http://cernvm.cern.ch/portal/
[6] http://www.opscode.com/chef/
[7] http://www.lustre.org
[8] http://xrootd.slac.stanford.edu/
[9] http://monalisa.caltech.edu/monalisa.htm
[10] http://www-panda.gsi.de/
[11] http://fairroot.gsi.de/
[12] http://alien2.cern.ch/
[13] http://serpiero.to.infn.it/
[14] http://cdash.gsi.de/CDash/index.php?project=PandaRoot
[15] http://www.fluka.org/fluka.php
[16] http://frankfurt-cloud.com/
[17] http://csc.uni-frankfurt.de/index.php?id=51
[18] http://pod.gsi.de/
[19] https://computing.llnl.gov/linux/slurm/
[20] http://www.fair-center.de
[21] http://www.gridka.de