Multi-Grid Job Scheduling

Klaus Rieger[*][†]
Astronomisches Rechen-Institut (ARI)
Zentrum für Astronomie (ZAH)
Universität Heidelberg
E-mail: rieger@ari.uni-heidelberg.de

Rainer Spurzem[‡]
National Astronomical Observatories of China (NAOC)
Chinese Academy of Sciences (CAS)
The Kavli Institut for Astronomy and Astrophysics
Peking University
E-mail: spurzem@bao.ac.cn

Astronomisches Rechen-Institut (ARI)
Zentrum für Astronomie (ZAH)
Universität Heidelberg
E-mail: spurzem@ari.uni-heidelberg.de

Combining computer resources from multiple administrative domains applied to a common task can be achieved using the grid technology. Grid Computing receives increasing attention in the astronomical community, where a great number of computer processing cycles and processing of large amounts of data are common. For example the astronomical virtual organizations in EGI and in OSG or Einstein@Home, discovering a pulsar using AstroGrid-D, based on Globus Toolkit and the metascheduler GridWay for job prioritization. To secure the availibilty of hosts and to obtain more information about the grid we have written AstroGridTest. Traditionally the available hosts of the community are limiting the possibilities of the scheduler. As part of the D-Grid Scheduler Interoperability project we break this limit and allow to send jobs even to hosts of other grid communities. Our main reason for the use of grids is optimized usage of the resources available on various hosts. Increasing performance on the path to exascale computing will at some stage require the power of an entire power plant if there is no fundamental change in architecture. On the way forward to Green Grid our work focuses on parallel processing use of power efficient GPU accelerators in grid environments. Here we describe how to build an international grid of clusters with such new architecture.

The International Symposium on Grids and Clouds (ISGC) 2012,
Febraury 26 – March 2, 2012
Academia Sinica, Taipei, Taiwan

[*]Speaker.

[†]This work is funded by the Deutsche Forschungsgemeinschaft (DFG) through SFB 881 “The Milky Way System” (Subproject Z2) at the Ruprecht-Karls-Universität Heidelberg.

[‡]Supported by the Chinese Academy of Sciences Visiting Professorship for Senior International Scientists,

1. Grid Computing

Grid Computing is widely used for astronomical calculations. For example the astronomical virtual organizations in EGI (27 VOs in this field) and in OSG (e.g. LIGO for the search of gravitational waves). The discovery of a pulsar in the data from the Arecibo Radio telescope, as another example, was achieved by increasing the processing power of Einstein@Home using AstroGrid-D. Grid technology allows combining computer resources from multiple administrative domains applied to a common task. Therefore computer grids are used to solve e.g. scientific problems, especially if they require a great number of processing cycles and to process large amounts of data.

1.1 Why one uses Grid Computing?

Traditionally the following points are mentioned for use of grid computing [1]:

- Improve efficiency
- Exploit idle resources
- Enable collaborations
- Create virtual resources and virtual organizations
- Increase capacity and productivity
- Parallel processing capacity
- Support different infrastructures
- Provide reliability and availability
- Reduce time to results
- Reduce costs

The main reason for the use of grids is the availability of various hosts. In addition aspects like parallel job processing and the use of power efficient methods get more and more important.

2. The community *AstroGrid-D*

Our group is part of the grid community *AstroGrid-D* [3][4][5]. The most important reasons to build up this grid are: We can easily use hosts of other members of the group (and vice versa), we can therefore get more computing power and last but not least we learn how to handle this interesting new technology to build up even more powerfull calculation communities in the future.

Most hosts of this community are located in Germany. However hosts of the *Main Astronomical Observatory of the National Academy of Sciences of Ukraine* and the *National Astronomical Observatories of China* are also accessible.

grant number 2009S1-5 ("The Silk Road Project"; Rainer Spurzem and Peter Berczik). The GPU accelerated computing clusters mentioned in this paper are the "`laohu`" cluster at the Center of Information and Computing at National Astronomical Observatories, Chinese Academy of Sciences, funded by Ministry of Finance of People's Republic of China under the grant ZDYZ2008-2, and the GRACE/GPU cluster "`titan`", funded under the grants I/80 041-043 and I/81 396 of the Volkswagen Foundation in Germany.

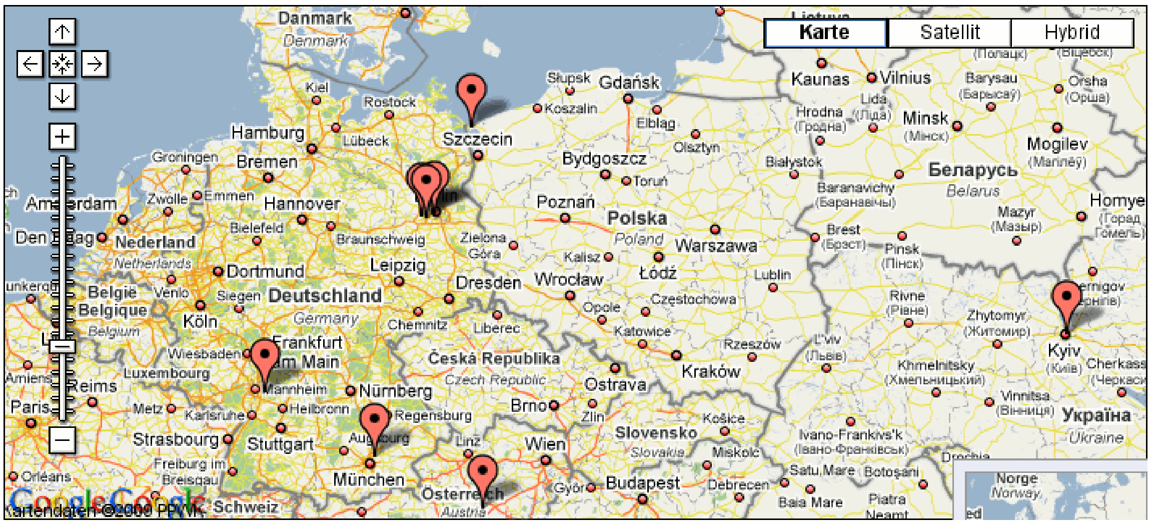


Figure 1: Locations of European hosts of the community *AstroGrid-D* [2]

2.1 Globus Toolkit

AstroGrid-D is based on *Globus Toolkit* [6], a toolkit for building computing grids developed and provided by the Globus Alliance [7].

2.2 Security

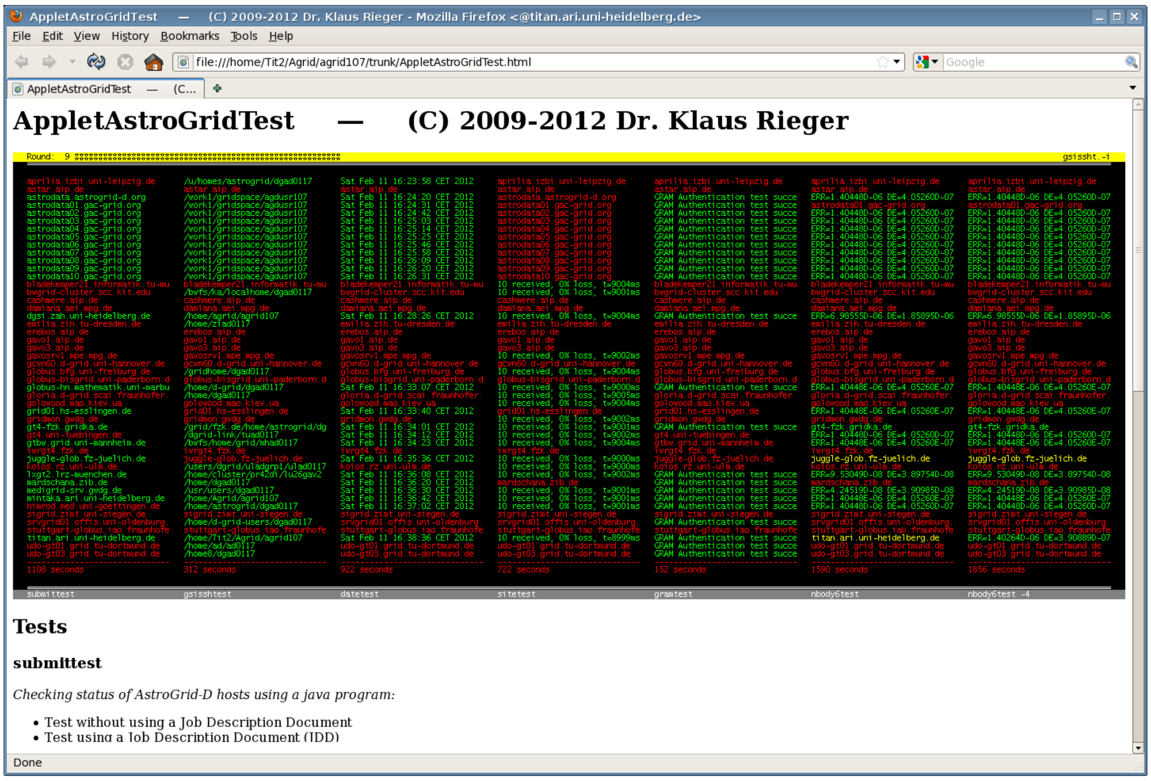
As an authentication mechanism the *Grid Security Infrastructure* (GSI)—based on the standard X.509—is used. Every user and service is identified by a certificate. Therefore, after creating the user credentials no password for login is required within the *AstroGrid-D* community.

2.3 Reliability

To secure the availibilty of hosts and to obtain more information about the grid *AstroGridTest* [8]—a compilation of several scripts and two kinds of graphical user interfaces—has been written using the script language Bash and the programming language Java. The scripts are checking the status of the hosts in various ways:

- Using a java program.
- Using a command for login (*gsissh*).
- Using bash commands like *date* or *sleep*.
- Using network checks like *ping* or *wsrf-query*.
- Using the Globus Resource Allocation Manager (GRAM) with *globusrun* or *globus-job-run*.
- Using simulation programs like *Nbody6++* [9][10].

The results are presented on the console or graphically with an independent program or with an applet running within a web browser (see Figure 2).



AppletAstroGridTest — (C) 2009-2012 Dr. Klaus Rieger - Mozilla Firefox <@titan.ari.uni-heidelberg.de>

File Edit View History Bookmarks Tools Help

file:///home/Tit2/Agrid/agrid107/trunk/AppletAstroGridTest.html

AppletAstroGridTest — (C) 2009-2012 Dr. Klaus Rieger

```
Round:  9 ***********************************************************                                                                                                                          gsissht.-i

aprilia.izbi.uni-leipzig.de      /u/homes/astrogrid/dgad0117      Sat Feb 11 16:23:58 CET 2012     aprilia.izbi.uni-leipzig.de      aprilia.izbi.uni-leipzig.de      aprilia.izbi.uni-leipzig.de      aprilia.izbi.uni-leipzig.de
astar.aip.de                     astar.aip.de                     astar.aip.de                     astar.aip.de                     astar.aip.de                     astar.aip.de                     astar.aip.de
astrodata.astrogrid-d.org        /work1/gridspace/agdusr107       Sat Feb 11 16:24:20 CET 2012     astrodata.astrogrid-d.org        GRAM Authentication test succe   ERR=1.40448D-06 DE=4.05260D-07   ERR=1.40448D-06 DE=4.05260D-07
astrodata01.gac-grid.org         /work1/gridspace/agdusr107       Sat Feb 11 16:24:31 CET 2012     astrodata01.gac-grid.org         GRAM Authentication test succe   astrodata01.gac-grid.org         astrodata01.gac-grid.org
astrodata02.gac-grid.org         /work1/gridspace/agdusr107       Sat Feb 11 16:24:42 CET 2012     astrodata02.gac-grid.org         GRAM Authentication test succe   ERR=1.40448D-06 DE=4.05260D-07   ERR=1.40448D-06 DE=4.05260D-07
astrodata03.gac-grid.org         /work1/gridspace/agdusr107       Sat Feb 11 16:25:03 CET 2012     astrodata03.gac-grid.org         GRAM Authentication test succe   ERR=1.40448D-06 DE=4.05260D-07   ERR=1.40448D-06 DE=4.05260D-07
astrodata04.gac-grid.org         /work1/gridspace/agdusr107       Sat Feb 11 16:25:14 CET 2012     astrodata04.gac-grid.org         GRAM Authentication test succe   ERR=1.40448D-06 DE=4.05260D-07   ERR=1.40448D-06 DE=4.05260D-07
astrodata05.gac-grid.org         /work1/gridspace/agdusr107       Sat Feb 11 16:25:25 CET 2012     astrodata05.gac-grid.org         GRAM Authentication test succe   ERR=1.40448D-06 DE=4.05260D-07   ERR=1.40448D-06 DE=4.05260D-07
astrodata06.gac-grid.org         /work1/gridspace/agdusr107       Sat Feb 11 16:25:46 CET 2012     astrodata06.gac-grid.org         GRAM Authentication test succe   ERR=1.40448D-06 DE=4.05260D-07   ERR=1.40448D-06 DE=4.05260D-07
astrodata07.gac-grid.org         /work1/gridspace/agdusr107       Sat Feb 11 16:25:58 CET 2012     astrodata07.gac-grid.org         GRAM Authentication test succe   ERR=1.40448D-06 DE=4.05260D-07   ERR=1.40448D-06 DE=4.05260D-07
astrodata08.gac-grid.org         /work1/gridspace/agdusr107       Sat Feb 11 16:26:09 CET 2012     astrodata08.gac-grid.org         GRAM Authentication test succe   ERR=1.40448D-06 DE=4.05260D-07   ERR=1.40448D-06 DE=4.05260D-07
astrodata09.gac-grid.org         /work1/gridspace/agdusr107       Sat Feb 11 16:26:20 CET 2012     astrodata09.gac-grid.org         GRAM Authentication test succe   ERR=1.40448D-06 DE=4.05260D-07   ERR=1.40448D-06 DE=4.05260D-07
astrodata10.gac-grid.org         /work1/gridspace/agdusr107       Sat Feb 11 16:26:31 CET 2012     astrodata10.gac-grid.org         GRAM Authentication test succe   ERR=1.40448D-06 DE=4.05260D-07   ERR=1.40448D-06 DE=4.05260D-07
bladekemper21.informatik.tu-mu   bladekemper21.informatik.tu-mu   bladekemper21.informatik.tu-mu   10 received, 0% loss, t=9004ms   bladekemper21.informatik.tu-mu   bladekemper21.informatik.tu-mu   bladekemper21.informatik.tu-mu
bwgrid-cluster.scc.kit.edu       /bwfs/ka/localhome/dgad0117      bwgrid-cluster.scc.kit.edu       10 received, 0% loss, t=9001ms   bwgrid-cluster.scc.kit.edu       bwgrid-cluster.scc.kit.edu       bwgrid-cluster.scc.kit.edu
cashmere.aip.de                  cashmere.aip.de                  cashmere.aip.de                  cashmere.aip.de                  cashmere.aip.de                  cashmere.aip.de                  cashmere.aip.de
damiana.aei.mpg.de               damiana.aei.mpg.de               damiana.aei.mpg.de               damiana.aei.mpg.de               damiana.aei.mpg.de               damiana.aei.mpg.de               damiana.aei.mpg.de
dgsi.zah.uni-heidelberg.de       /home/agrid/agrid107             Sat Feb 11 16:28:26 CET 2012     10 received, 0% loss, t=9004ms   GRAM Authentication test succe   ERR=6.98555D-06 DE=1.85895D-06   ERR=6.98555D-06 DE=1.85895D-06
emilia.zih.tu-dresden.de         /home/ziad0117                   emilia.zih.tu-dresden.de         emilia.zih.tu-dresden.de         emilia.zih.tu-dresden.de         emilia.zih.tu-dresden.de         emilia.zih.tu-dresden.de
erebos.aip.de                    erebos.aip.de                    erebos.aip.de                    erebos.aip.de                    erebos.aip.de                    erebos.aip.de                    erebos.aip.de
gavo1.aip.de                     gavo1.aip.de                     gavo1.aip.de                     gavo1.aip.de                     gavo1.aip.de                     gavo1.aip.de                     gavo1.aip.de
gavo3.aip.de                     gavo3.aip.de                     gavo3.aip.de                     gavo3.aip.de                     gavo3.aip.de                     gavo3.aip.de                     gavo3.aip.de
gavosrv1.mpe.mpg.de              gavosrv1.mpe.mpg.de              gavosrv1.mpe.mpg.de              10 received, 0% loss, t=9002ms   gavosrv1.mpe.mpg.de              gavosrv1.mpe.mpg.de              gavosrv1.mpe.mpg.de
gcwn60.d-grid.uni-hannover.de    gcwn60.d-grid.uni-hannover.de    gcwn60.d-grid.uni-hannover.de    gcwn60.d-grid.uni-hannover.de    gcwn60.d-grid.uni-hannover.de    gcwn60.d-grid.uni-hannover.de    gcwn60.d-grid.uni-hannover.de
globus.bfg.uni-freiburg.de       /gridhome/dgad0117               globus.bfg.uni-freiburg.de       10 received, 0% loss, t=9004ms   globus.bfg.uni-freiburg.de       globus.bfg.uni-freiburg.de       globus.bfg.uni-freiburg.de
globus-bisgrid.uni-paderborn.d   globus-bisgrid.uni-paderborn.d   globus-bisgrid.uni-paderborn.d   globus-bisgrid.uni-paderborn.d   globus-bisgrid.uni-paderborn.d   globus-bisgrid.uni-paderborn.d   globus-bisgrid.uni-paderborn.d
globus-hn.mathematik.uni-marbu   /home/d-grid/dgad0117            Sat Feb 11 16:33:07 CET 2012     10 received, 0% loss, t=9000ms   GRAM Authentication test succe   ERR=1.40448E-06 DE=4.05260E-07   ERR=1.40448E-06 DE=4.05260E-07
gloria.d-grid.scai.fraunhofer.   /home/dgad0117                   gloria.d-grid.scai.fraunhofer.   10 received, 0% loss, t=9005ms   gloria.d-grid.scai.fraunhofer.   gloria.d-grid.scai.fraunhofer.   gloria.d-grid.scai.fraunhofer.
golowood.mao.kiev.ua             golowood.mao.kiev.ua             golowood.mao.kiev.ua             10 received, 0% loss, t=9004ms   golowood.mao.kiev.ua             golowood.mao.kiev.ua             golowood.mao.kiev.ua
grid01.hs-esslingen.de           grid01.hs-esslingen.de           Sat Feb 11 16:33:40 CET 2012     grid01.hs-esslingen.de           grid01.hs-esslingen.de           ERR=1.40448E-06 DE=4.05260E-07   ERR=1.40448E-06 DE=4.05260E-07
gridmon.gwdg.de                  gridmon.gwdg.de                  gridmon.gwdg.de                  10 received, 0% loss, t=9002ms   gridmon.gwdg.de                  gridmon.gwdg.de                  gridmon.gwdg.de
gt4-fzk.gridka.de                /grid/fzk.de/home/astrogrid/dg   Sat Feb 11 16:34:01 CET 2012     10 received, 0% loss, t=9001ms   GRAM Authentication test succe   gt4-fzk.gridka.de                gt4-fzk.gridka.de
gt4.uni-tuebingen.de             /dgrid-link/tuad0117             Sat Feb 11 16:34:12 CET 2012     10 received, 0% loss, t=9002ms   gt4.uni-tuebingen.de             ERR=1.40448D-06 DE=4.05260D-07   ERR=1.40448D-06 DE=4.05260D-07
gtbw.grid.uni-mannheim.de        /bwfs/home/grid/mhad0117         Sat Feb 11 16:34:23 CET 2012     10 received, 0% loss, t=9004ms   gtbw.grid.uni-mannheim.de        ERR=1.40448E-06 DE=4.05260E-07   ERR=1.40448E-06 DE=4.05260E-07
iwrgt4.fzk.de                    iwrgt4.fzk.de                    iwrgt4.fzk.de                    iwrgt4.fzk.de                    iwrgt4.fzk.de                    iwrgt4.fzk.de                    iwrgt4.fzk.de
juggle-glob.fz-juelich.de        juggle-glob.fz-juelich.de        Sat Feb 11 16:35:36 CET 2012     10 received, 0% loss, t=9000ms   juggle-glob.fz-juelich.de        juggle-glob.fz-juelich.de        juggle-glob.fz-juelich.de
koios.rz.uni-ulm.de              /users/dgrid/uladgrp1/ulad0117   koios.rz.uni-ulm.de              10 received, 0% loss, t=9000ms   koios.rz.uni-ulm.de              koios.rz.uni-ulm.de              koios.rz.uni-ulm.de
lxgt2.lrz-muenchen.de            /home/cluster/pr42di/lu26gav2    Sat Feb 11 16:36:08 CET 2012     10 received, 0% loss, t=9002ms   GRAM Authentication test succe   ERR=9.53049D-08 DE=3.89754D-08   ERR=9.53049D-08 DE=3.89754D-08
mardschana.zib.de                /home/dgad0117                   Sat Feb 11 16:36:20 CET 2012     mardschana.zib.de                GRAM Authentication test succe   mardschana.zib.de                mardschana.zib.de
medigrid-srv.gwdg.de             /usr/users/dgad0117              Sat Feb 11 16:36:30 CET 2012     10 received, 0% loss, t=9001ms   GRAM Authentication test succe   ERR=4.24519D-08 DE=3.90985D-08   ERR=4.24519D-08 DE=3.90985D-08
mintaka.ari.uni-heidelberg.de    /home/Agrid/agrid107             Sat Feb 11 16:36:42 CET 2012     10 received, 0% loss, t=9001ms   GRAM Authentication test succe   ERR=1.40448E-06 DE=4.05260E-07   ERR=1.40448E-06 DE=4.05260E-07
nimrod.med.uni-goettingen.de     /home/astrogrid/dgad0117         Sat Feb 11 16:37:02 CET 2012     10 received, 0% loss, t=9001ms   GRAM Authentication test succe   ERR=1.40448D-06 DE=4.05260D-07   ERR=1.40448D-06 DE=4.05260D-07
sigrid.zimt.uni-siegen.de        sigrid.zimt.uni-siegen.de        sigrid.zimt.uni-siegen.de        sigrid.zimt.uni-siegen.de        sigrid.zimt.uni-siegen.de        sigrid.zimt.uni-siegen.de        sigrid.zimt.uni-siegen.de
srvgrid01.offis.uni-oldenburg.   /home/d-grid-users/dgad0117      srvgrid01.offis.uni-oldenburg.   srvgrid01.offis.uni-oldenburg.   GRAM Authentication test succe   srvgrid01.offis.uni-oldenburg.   srvgrid01.offis.uni-oldenburg.
stuttgart-globus.iao.fraunhofe   stuttgart-globus.iao.fraunhofe   stuttgart-globus.iao.fraunhofe   stuttgart-globus.iao.fraunhofe   stuttgart-globus.iao.fraunhofe   stuttgart-globus.iao.fraunhofe   stuttgart-globus.iao.fraunhofe
titan.ari.uni-heidelberg.de      /home/Tit2/Agrid/agrid107        Sat Feb 11 16:38:36 CET 2012     10 received, 0% loss, t=8999ms   GRAM Authentication test succe   titan.ari.uni-heidelberg.de      ERR=1.40264D-06 DE=3.90889D-07
udo-gt01.grid.tu-dortmund.de     /home/ad/ad0117                  udo-gt01.grid.tu-dortmund.de     udo-gt01.grid.tu-dortmund.de     GRAM Authentication test succe   udo-gt01.grid.tu-dortmund.de     udo-gt01.grid.tu-dortmund.de
udo-gt03.grid.tu-dortmund.de     /home8/dgad0117                  udo-gt03.grid.tu-dortmund.de     udo-gt03.grid.tu-dortmund.de     GRAM Authentication test succe   udo-gt03.grid.tu-dortmund.de     udo-gt03.grid.tu-dortmund.de
------------------------------   ------------------------------   ------------------------------   ------------------------------   ------------------------------   ------------------------------   ------------------------------
1108 seconds                     312 seconds                      922 seconds                      722 seconds                      152 seconds                      1590 seconds                     1856 seconds

submittest                       gsisshtest                       datetest                         sitetest                         gramtest                         nbody6test                       nbody6test -4
```

Tests

submittest

Checking status of AstroGrid-D hosts using a java program:

- Test without using a Job Description Document
- Test using a Job Description Document (JDD)

Done

Figure 2: *AppletAstroGridTest*: Clear presentation of several tests of hosts of the *AstroGrid-D* community presented within a web browser. *JavaAstroGridTest* is a similar, however stand-alone running program.

2.4 Data Management

The *AstroGrid-D Data Management* (ADM) [11] provides a virtual filesystem and simultaneously cares for the placement of the corresponding physical files on one or more storage facilities. It is a tool for distributed data-management.

ADM uses a relational database to store a unique descriptor as well as meaningful or typical meta data for each file or directory, e.g. the owner and a timestamp to log when the entry has been registered within the filesystem. ADM provides a command line client and a web interface allowing to browse the virtual filesystem graphically.

2.5 Discovery, Negotiation and Scheduling

It is important to get information about the hosts of the community in order to choose the best one available. Three sources provide this information: First, *MDS* [12][13], the monitoring and discovery system, gets data from the host, from *Ganglia* and from *GeoMaint*. Second, *Ganglia* [14][15], a monitoring system for high-performance computing systems provides information about the host and—via net—about other hosts. Third there is *GeoMaint* [16][17], an information-provider (sensor) for geolocation and maintenance data.

The available information allows a good decision either for the user or for the automatic system, which is described in the following section.

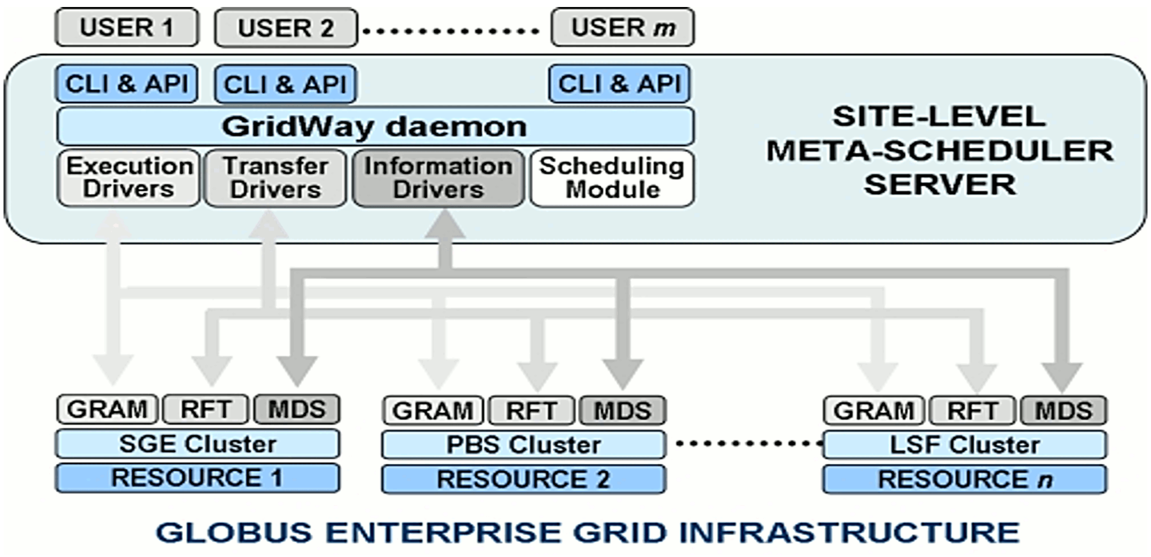


Figure 3: The Metascheduler *GridWay* seperates the user layer (various users with CLI and API) from the lower layers (Middleware Access Drivers: GRAM (or other LRM), RFT and MDS). Source: www.globus.org/toolkit/docs/4.0/contributions/gridway/admin-index.html

2.6 Negotiation and Scheduling

The Metascheduler *GridWay* [18][19][20] uses two policies for prioritization: The job and the resource prioritization. For example the waiting time is part of the job, the necessary requirements are part of the resource. This prioritization allows *GridWay* to decide automatically which host fits best for the job execution.

An even further step is the aim of the *D-Grid Scheduler Interoperability* (DGSI) project [21]: Traditionally the available hosts of the community are limiting the possibilities of the scheduler. The DGSI project breaks this limit and allows to send jobs even to hosts of other grid communities. To achieve these goals it is necessary to connect grids to an interoperability layer by using public standards like the service level agreement layer *SLA4D-Grid* [22] as far as possible.

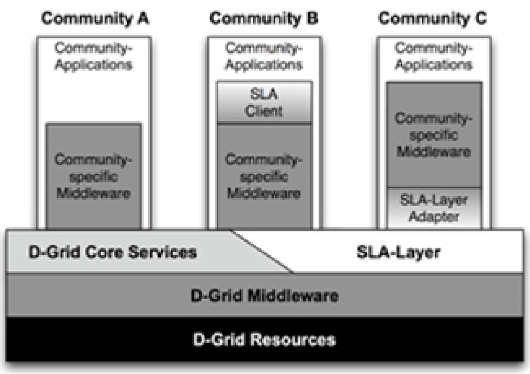


Figure 4: *SLA4D-Grid*, a service level agreement layer for the D-Grid placed between the D-Grid middleware (e.g. *Globus Toolkit*) and the D-Grid user layer. Source: http://www2.fz-juelich.de/jsc/grid/sla4d-grid

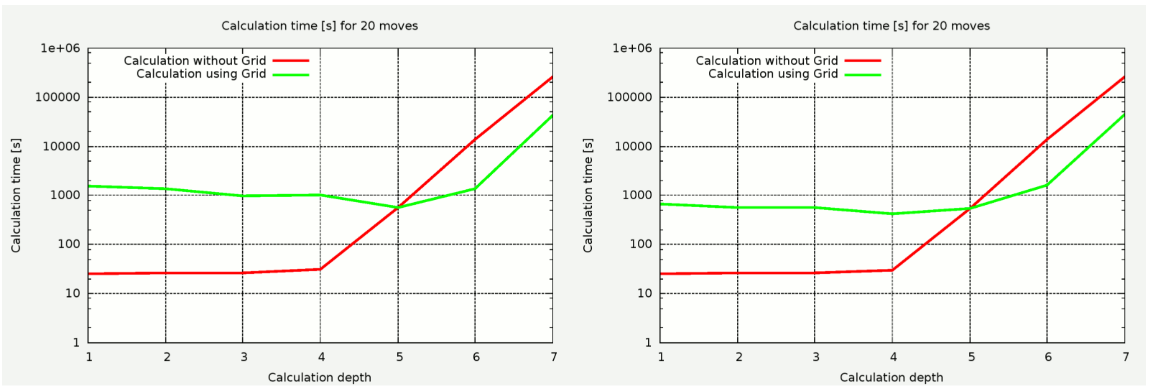


Figure 5: *Hosts used for the first test (diagram left side)*: astar.aip.de, astrodata.astrogrid-d.org, astrodata01.gac-grid.org, . . . , astrodata10.gac-grid.org, cashmere.aip.de, grid01.hs-esslingen.de, gt4-fzk.gridka.de, gt4.uni-tuebingen.de, iwrgt4.fzk.de, mintaka.ari.uni-heidelberg.de. *Hosts used for the second test (diagram right side)*: astar.aip.de, astrodata.astrogrid-d.org, astrodata01.gac-grid.org, . . . , astrodata10.gac-grid.org, cashmere.aip.de, dgsi.zah.uni-heidelberg.de, gt4-fzk.gridka.de, gt4.uni-tuebingen.de, iwrgt4.fzk.de, mintaka.ari.uni-heidelberg.de, titan.ari.uni-heidelberg.de.

3. Use cases

3.1 Use best host for astronomical simulation Nbody6++

The deployment package *nb6deployment* [23] allows the download of the astronomical simulation software *Nbody6++* and its submission to a host of the grid in an easy way. It is possible to run the program on a distinguished host or to use GridWay for choosing the best available host automatically.

3.2 Using hosts in parallel for artificial intelligence problem

The following use case shows if there is a speed-up by distributing a real world algorithm over the grid. As an example for artificial intelligence the game *Nine Men's Morris* [24] has been written in Java. The solving algorithm depends on two important settings:

1. *Depth* determines how many steps the algorithm considers in advance. Greater depth can result in better predictions and therefore better strength, however, much more computation time is needed.

2. *Grid* determines, if the calculation after the first step gets divided and then submitted to various hosts of the grid. Therefore, it is possible to compare the time needed with or without using the hosts of the community.

Test runs: Two test runs with slightly different hosts have been conducted (see Figure 5).

Results and Conclusion: The results are similar: In case of *depth* between one and four using the grid leads to a slower calculation because there is a lot of overhead necessary to distribute the jobs and to collect the results later. Therefore, in the case of simple calculations it is faster not to use grid computing.

However, for *depth* greater than five grid computing is almost one order of magnitude faster. It should be noted that optimization of the distribution algorithm is expected to result in further performance enhancement. Therefore, in case of time consuming calculations the use of various hosts in parallel can reduce the time effort dramatically.

Similar speed enhancements for astronomical problems are possible, if calculations can run separately as they do for the artificial intelligence problem.

4. Green Grid

While in the past the advantage of having more available hosts was mainly to get more computing power by choosing the most powerful host available or combining the power of several hosts, there are new challenges for the future: How can we reduce the power consumption?

4.1 Strongly growing power consumption

Computing performance has grown and so has the power consumption. Any increase in power consumption also means much higher cooling demands. This point is very important because approximately 75 % of the amount of electrical energy used for computing is required just to run the cooling system. Increasing performance on the path to exascale computing will at some stage require the power of an entire power plant if there is no fundamental change in architecture.

In addition, comparing the performance of test cases with the performance of real applications shows an decreasing efficiency for years (see Figure 6).

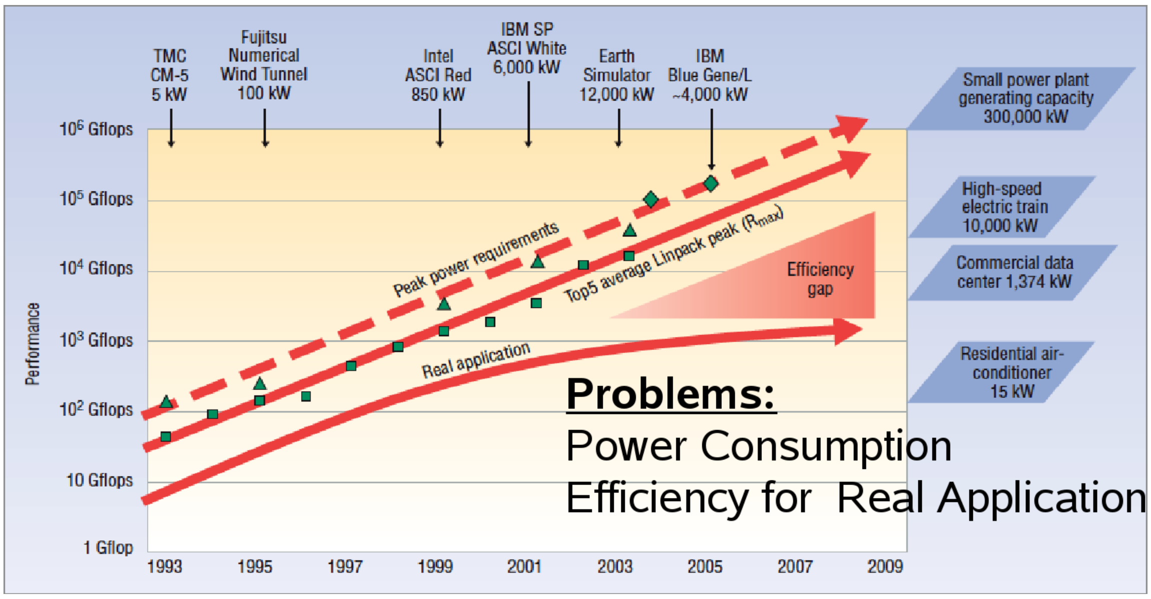


Figure 6: Computing performance has grown and so has the power consumption. However, since the efficiency gap gets larger the benefit for real applications is lower than expected according to tests. Chart:[25]

4.2 Saving power using special-purpose processors

Clusters with specialized hardware are reaching high performance with less power consumption. Following the special-purpose processors *GRAPE*, *GPU* and *FPGA* are described.

4.2.1 GRAPE

During the time of introduction this processor was the fastest solution available. There is a lot of experience because this processor was successfully used for astronomical calculations for many years. However, it is not a real power saver and it is not developed any further. In a nutshell: Although not the future, it is still suitable for some physical simulations.

At ARI/ZAH we are using specialized processors (phiGRAPE) since many years for solving N-Body6 problems (see Figure 7). Although these processors are not developed any further N-Body6 remains a suitable testcase because such physical simulations demand steadily increasing computational power.

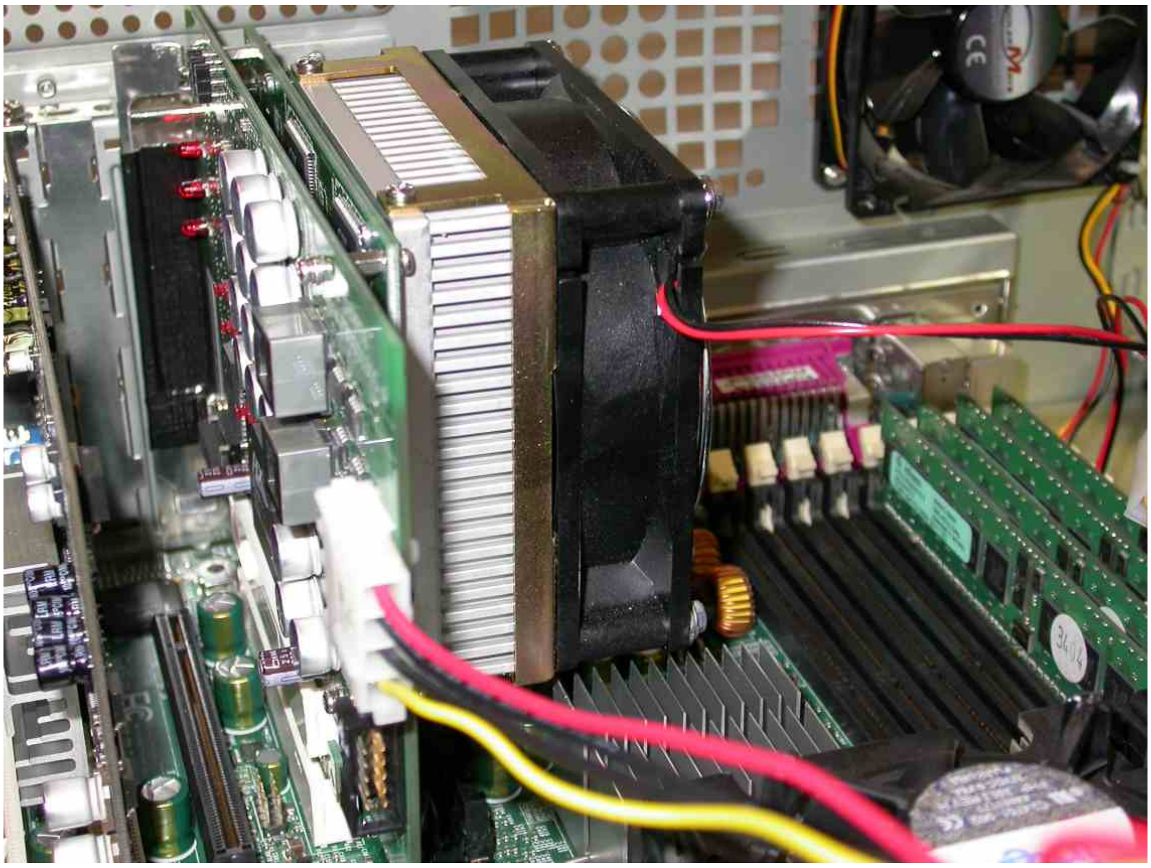


Figure 7: Cards with *GRAPE* running successfully in our host "titan" for years. We are using the specialized processors *phiGRAPE* for solving N-Body6 problems.

4.2.2 GPU

*G*raphics *P*rocessing *U*nits are originally intended for manipulating computer graphics. Nevertheless their highly parallel structure makes them more effective than general-purpose CPUs for a range of complex algorithms. Due to the mass market, they achieve high performance at an affordable price. In relation to their performance they save electricity (see Figure 8).

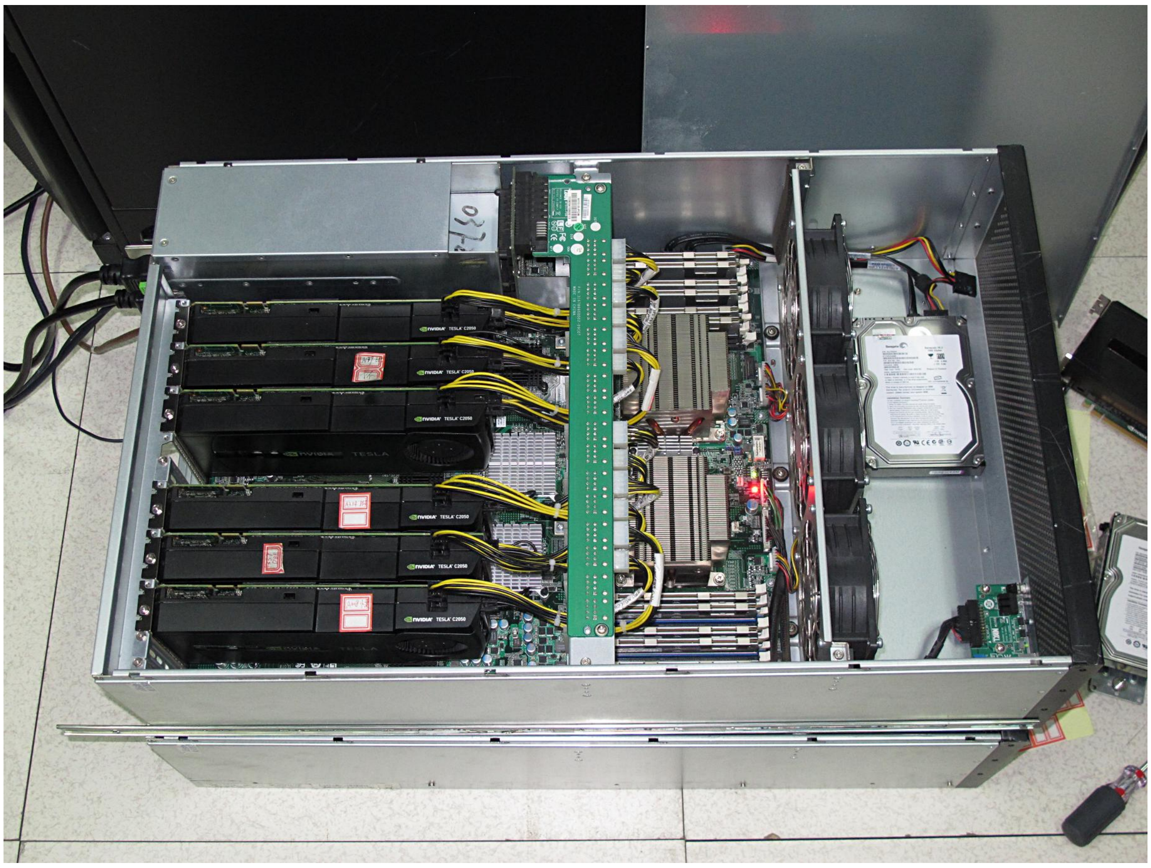


Figure 8: Six *GPU* cards on one board used for the Mole-8.5 supercomputer.

4.2.3 FPGA

*F*ield-*P*rogrammable *G*ate *A*rrays contain programmable logic components and a hierarchy of reconfigurable interconnects. They are intended for applications making use of massive parallelism offered by their architecture.

The programming in VHDL (*V*ery High Speed Integrated Circuits *H*ardware *D*escription *L*anguage) is difficult and programming is a complex multi-step and therefore time-consuming process.

FPGAs are getting cheaper while capacity and computing power are increasing. Advantages of the FPGAs are the small efficiency gap and the low power consumption.

5. Saving power choosing the most efficient host

The availability of various new architectures allows to perform calculations in a more effective way. Both GPUs and FPGAs are attractive for certain applications. On the one hand programming FPGAs demands a lot of effort to save power using an optimized code, on the other hand as basic elements of architecture they are able to save more power then GPUs. This is the reason why we—in parallel to the support of FPGA boards—are in the process of running more and more clusters of GPUs in Germany and China.

In addition to CPUs both special-purpose processors are of interest. The job determines the most suitable host, because calculation speed and power consumption is job dependent. There is no one-processor-fits-all solution. Therefore, it is necessary to provide several kinds of computer architectures and to enhance schedulers (like GridWay) in order to find the *best fitting host*—the host performing high computational power while saving energy.

Figure 9: Computational power in a high energy-efficient way using *FPGA*: *MP RACE-2* (Virtex 4, XC4FX60 PCIe interface x4 (1GB/s), 16 MB DDR SRAM, DDR SDRAM SO-DIMM slot) [26]

References

[1] L. Ferreira et al., *Grid Computing in Research and Education*, IBM Redbooks, April 2005.

[2] LRZ München, *Dynamically created resource provider card*, http://webmds.lrz.de:8080/webmds/xslfiles/csm/.

[3] H. Enke et al., *AstroGrid-D: Grid Technology for Astronomical Science*, *New Astron.* 16, 79–93, 2011 [1007.4053v1].

[4] Th. Brüsemeister, *First steps in AstroGrid-D: Chapter 1 — First steps in AstroGrid-D*, http://www.ari.uni-heidelberg.de/mitarbeiter/bruesemeister/agdguide/c11.htm.

[5] T. Beck-Ratzka, R. Spurzem, K. Rieger, Th. Brüsemeister et al., *AstroGrid-D: Grid User Guide — Guide for porting User Applications to the Grid*, http://www.gac-grid.de/project-documents/Posters/App2Grid-Guide-AstroGrid-1.1.pdf.

[6] K. Rieger, *Installation of Globus Toolkit 4.0.8 on Scientific Linux 5.5 running virtually within VMware on openSUSE 11.1*, http://www.ari.uni-heidelberg.de/mitarbeiter/rieger/InstallationGlobusToolkitScientificLinux.html.

[7] Globus Alliance, *Globus Toolkit*, http://globus.org/toolkit/.

[8] K. Rieger, *AstroGridTest*, Manual: http://www.ari.uni-heidelberg.de/mitarbeiter/rieger/manual.pdf, Software: svn://svn.gac-grid.org/software/AstroGridTest/trunk/.

[9] A. Borch, R. Spurzem, J. Hurley, *NBODY meets stellar population synthesis*, 2005 [0704.3915v1].

[10] E. Khalisi, P. Amaro-Seoane, R. Spurzem, *A comprehensive NBODY study of mass segregation in star clusters: energy equipartition and escape*, 2007 [astro-ph/0602570v2].

[11] Th. Brüsemeister, *AstroGrid-D Data Management — Tutorial*, http://www.gac-grid.de/project-products/Software/adm/adm-tutorial.pdf.

[12] Globus Alliance, *GT Information Services: Monitoring & Discovery System (MDS)*, http://www.globus.org/toolkit/mds/.

[13] K. Rieger, *Use WebMDS*, Use WebMDS.

[14] *Ganglia Monitoring System*, http://ganglia.info.

[15] K. Rieger, *Install and Use Ganglia*, Install Ganglia, Use Ganglia.

[16] *GeoMaint Informationprovider for Geolocation and Maintenance data*, http://sourceforge.net/projects/geomaint/.

[17] K. Rieger, *Setup of Information Provider MDS4 and GeoMaint*, http://www.ari.uni-heidelberg.de/mitarbeiter/rieger/SetupGeoMaint.html.

[18] *GridWay Metascheduler*, http://gridway.org.

[19] K. Rieger, *Installation of GridWay and GridGateWay*, Installation of GridWay 5.6.1, Installation of GridWay 5.8.1 and GridGateWay 1.0.4.

[20] K. Rieger, *Using GridWay*, http://www.ari.uni-heidelberg.de/mitarbeiter/rieger/UsingGridWay.html.

[21] The D-Grid Scheduler Interoperability (DGSI) project, http://dgsi.d-grid.de/.

[22] *SLA4D-Grid*, http://www2.fz-juelich.de/jsc/grid/sla4d-grid/.

[23] Th. Brüsemeister, *First steps in AstroGrid-D: Chapter 2 — NBODY6++ Deployment*, Manual: http://www.ari.uni-heidelberg.de/mitarbeiter/bruesemeister/agdguide/c93.htm, Software: http://svn.ari.uni-heidelberg.de/repos/nbody/deployment/branches/0.2.x/.

[24] K. Rieger, *Grid Nine Men's Morris*, http://www.ari.uni-heidelberg.de/mitarbeiter/rieger/GridNineMensMorris.html.

[25] H. D. Simon, *Lawrence Berkeley National Laboratory, USA*, 2005.

[26] R. Spurzem, P. Berczik, G. Marcus, et al., *Accelerating Astrophysical Particle Simulations with Programmable Hardware (FPGA and GPU)*, *Computer Science — Research and Development* Volume 23, Numbers 3–4, 231–239.