GES 2007 German e-Science

Available online at http://www.ges2007.de

Accounting Facilities in the European Supercomputing Grid DEISA

Johannes Reetz[1], Thomas Soddemann[1], Bart Heupers[2], Jules Wolfrat[2]

[1]Garching Computing Centre of the Max Planck Society
Max-Planck-Institute for Plasma Physics
D-85748 Garching, Germany
Email: {johannes.reetz, thomas.soddemann}@rzg.mpg.de
Phone: (+49 89) 3299 2199, *fax:* (+49 89) 3299 1301

[2] SARA, Computing and Networking Services
HPC department
Kruislaan 415, 1090 GP Amsterdam, The Netherlands
Email: {bart,wolfrat}@sara.nl

Abstract

Account management and resource usage monitoring are essential services for production Grids. The scope of a production Grid infrastructure, the heterogeneity of resources and services, the typical community usage profiles, and the depth of integration of the resource providers regarding operational procedures and policies imply specific requirements for accounting facilities. We present the accounting facilities currently used in production in the Distributed European Infrastructure for the Supercomputing Applications (DEISA). DEISA is a consortium of leading national supercomputing centres currently deploying and operating a persistent, production quality, distributed supercomputing environment with continental scope. The DEISA accounting facilities gather information from the site-local batch systems and the distributed DEISA user administration system, and generate XML usage records conforming to the OGF usage record specification which are then stored locally in a XML data base at each DEISA site. The distributed accounting information can be fetched by clients such as users, project supervisors, site accounting managers and DEISA supervisors. The information is made available by site-local WSRF-compliant accounting information services that allow for a fine-grained setting of access rights. Each authorized client gets a specific view on the accounting information according to one of the following roles: a) a *site accounting manager* imports usage records of related home-site users from all DEISA sites for long-term archiving, b) a *project supervisor* retrieves information to assess the resource usage by his project partners, c) a *DEISA supervisor* (e.g. someone overlooking the usage on behalf of the DEISA executive committee) gets a report on the global usage of DEISA resources, and d) the user who can retrieve all the accounting information related to his own jobs. The privacy and integrity of the data provided and transferred from the accounting information service running at each site is guaranteed using X.509 certificates for mutual authentication and secure communication channels.

1 Introduction

The monitoring of the distributed resource usage is essential in production Grids. This information allows billing users on a *pay-per-use* basis or balancing the resource usage between the user home sites collaborating in a Grid infrastructure. The former represents the general business case in Grid computing and is targeted by many accounting systems on the market. DEISA[1] is interested to get the resource usage information for balancing the resource usage between the partner sites and for checking if projects are not exceeding their budgets. In the following we report the accounting facilities used by DEISA. Different from an integrated accounting system, these facilities are building blocks for exchanging resource usage information between the DEISA partners. So, they represent lightweight easy-to-deploy facilities rather than a complete system.

1.1 DEISA - the European Supercomputing Grid

DEISA, the Distributed European Infrastructure for Supercomputing Applications, is a consortium of leading national supercomputing centres funded by the 6th European Framework Program.

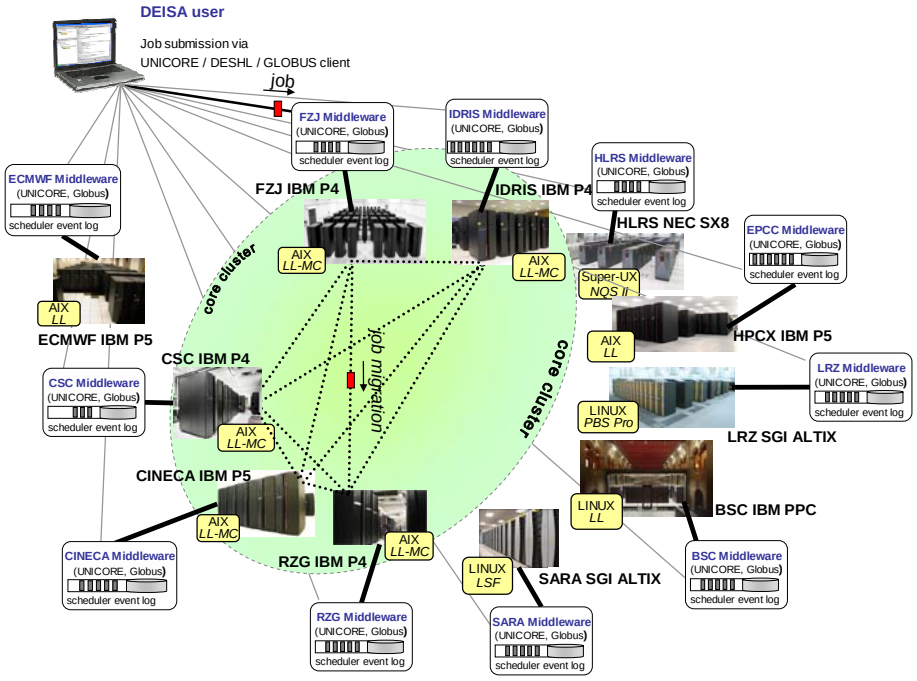

Figure 1: The DEISA Grid infrastructure. The dotted lines connect those sites employing the multi-cluster LoadLeveler (LL-MC) that allows for job migration without any middleware involved. As a use case, a DEISA user submits a job via UNICORE to FZJ. The LL-MC at FZJ migrates the job to RZG. Generally, jobs can be submitted either directly from a login node at the user's home site or using the UNICORE infrastructure [2]. In future, jobs can also be submitted via Globus[3].

The deployed production-quality supercomputing Grid infrastructure is based on a deep integration of existing high-performance computing platforms tightly coupled by a dedicated network, a transparent wide-area multi-

platform shared file system and innovative Grid middleware such as UNICORE. The deep integration of the DEISA partner sites is reflected by the collaborative management of DEISA user accounts (see section 2.1).

Eleven DEISA partners provide high-performance computer systems equipped with various batch schedulers running under different operating systems (Fig.1): CINECA/Italy, CSC/Finland, FZJ/Germany, IDRIS/France, and RZG/Germany provide IBM P4 and P5 machines with the multi-cluster LoadLeveler (LL-MC) running under AIX. Note that the LL-MC enables direct job-rerouting between sites without any middleware involved in between. The partners ECWMF and EPCC/UK provide also IBM P5 systems but currently with LoadLeveler (LL) running under AIX. LRZ/Germany and SARA/Netherlands provide SGI Altix machines running Linux, LRZ employs the batch scheduler PBS Pro while SARA uses LSF. BSC/Spain provide a Power PC cluster running currently LL under Linux up to now, and the HLRS/Germany provides a NEC SX-8 running the NQS-II batch system under Super-UX. So, there is a variety of batch schedulers and operating systems with an emphasis on the IBM LoadLeveler. The job migration mechanism within the core cluster is relevant for gathering resource usage information because getting the data from the scheduler event log-files of any Grid middleware is not appropriate in this case (section 2.3).

1.2 Existing Accounting Systems for Grid projects

Several accounting systems have been developed and are actually in production use in Grid projects. For a detailed review of the various accounting systems see [4]. In early 2006, DEISA evaluated some of the available accounting systems (APEL[5], DGAS[6], SGAS[7] and AMIE[8]).

SGAS is certainly one of the most elaborate representatives in the realm of accounting systems. It offers a bank service, a job account reservation manager (JARM), and a log and usage tracking service (LUTS). The bank service is typically located at a virtual organization and holds the user's or organization's funds. The JARM service allocates funds from a given bank account when a job is submitted to a resource manager (Globus WS-GRAM). After completion of a job the accounting information are forwarded to the LUTS and the equivalent of the resource usage in funds is finally withdrawn from the bank account and unused funds are freed. Unfortunately SGAS has a completely Globus Toolkit centric implementation.

DGAS is an accounting system used by the EGEE Grid project. It follows a decentralized approach in storing accounting information. Resource usage information from a user job at site *A* is transferred to the home site of this user. So, home-sites can track the resource consumption of their users. DGAS offers a policy enforcement point in combination with the DataGrid Workload Manager. Unfortunately the system lacks the aspect of directly accounting whole sites e.g. in order to support a zero balance resource exchange between a collaboration of sites. Furthermore, in 2006, DGAS did not employ standards such as the OGF usage record format.

APEL is an accounting and billing software that is currently used within the EGEE Grid project. It supports processing local resource management

systems' log-files for gathering accounting data. These data are transmitted to a central database from where it can be retrieved by authorized parties. While the original usage records remain stored locally at the job execution site, a duplicate of each of them is replicated into the central database.

TeraGrid's AMIE is an integrated user management and accounting system. It comprises a central database which contains information on users as well as accounting information. User accounts at a job execution site in TeraGrid are created according to the account information provided by AMIE's database. After a user's job is finished the accounting data are reported back to AMIE's central database. AMIE then can subtract the consumed computing time from the previously granted time. AMIE does not (yet) support open formats or follow an open interface specification.

2 Analysis of DEISA Requirements for Accounting

The DEISA requirements for accounting are motivated by the project aim to enable challenging scientific computations being executed on high-end super-computing clusters providing high-bandwidth and low-latency interconnects. This implies that typically large DEISA compute jobs are running over a longer period of time. These jobs are submitted to a cluster batch system, preferentially via Grid middleware. It must be considered that jobs can migrate from one site to another, directly dispatched by the batch scheduler of a partner site (Fig.1). Furthermore, interactive logins of users from remote home sites are disabled in general, so that any interactive usage of compute facilities should be negligible. Thus, it is sufficient to determine the resource usage information exclusively from the data in the batch system log-files.

2.1 The User-Administration System (UAS)

The DEISA user administration system is an important source of information for accounting. It is based on a distributed network of LDAP servers and a set of agreed operational and administrative policies. End-to-End security is ensured using SSL and X.509 server certificates for mutual authentication and authorisation of the LDAP clients and servers. Each site operates an own LDAP server publishing account information exclusively concerning their assigned DEISA users. The site that registers the user is referred to as the *home site* of the user.

Every site imports the user account information from the other LDAP servers periodically, at least daily, and the corresponding DEISA user accounts are created or modified by the local account management systems. A DEISA user account remains valid within DEISA until the corresponding user entry is disabled by the administrator at the user's *home site*. Specific uid-ranges for each DEISA site and a naming convention for usernames ensure that the user accounts remain unique within DEISA. Additionally, every user must be assigned to one project. The UAS provides also information about projects and the project membership. Partner sites may replicate the data from the UAS to local directory servers or local data bases. In any case the maximum synchronisation delay time for the UAS is one day.

2.2 Constraints

A requirement catalogue must consider the following constraints that can not be translated directly into requirements:

1. DEISA jobs are scheduled and executed in batch mode
2. LL-MC (LL) is currently the most employed batch scheduler in DEISA
3. jobs can migrate between sites at system resource management level
4. each user is assigned to a home site that manages his DEISA account
5. each user is assigned to a project supervised by a project responsible who also is dedicated to a DEISA home site
6. DEISA follows the zero balance resource exchange strategy: each site gives a certain percentage of its compute resource into a resource pool
7. the DEISA Executive Committee needs regular reports on the status and the progress of the resource consumption of all the projects in DEISA
8. the compute resources in DEISA comprise mainly clusters of multi-processor nodes (parallel computers); Simultaneous sharing of nodes between jobs must be avoided to prevent a degradation of performance
9. the DEISA UAS is the primary source of information about users and projects; the maximum synchronisation interval of the system (e.g., user records are modified at his home site) is one day
10. local directories/databases mirror the DEISA user and project information locally and may be organized and operated specifically at each site
11. the usage records gathered at each site are sensitive information so that any exchange of accounting information must follow legal rules (EU directive 95/46/EC [9] on the protection of individuals regarding the processing of personal data and on the free movement of such data

2.3 Requirements

With regard to the constraints presented above the requirements are:

1. The facilities for gathering usage information from the resource management log-file must be easy to deploy and reliable. The batch schedulers employed at the DEISA sites must be supported in particular.
2. The accounting facilities must not depend on a specific middleware.
3. Avoiding a central repository that stores all the sensitive accounting information in detail, the use of secured local databases is preferred.
4. The local accounting databases must be synchronised periodically (daily) with combined information from the scheduler logs and the UAS.
5. The accounting database needs an identity- and role-based access control.
6. Role-specific views on the accounting information: specific retrieved data may be complete for the roles *user* and (home) *site accounting manager* (gets only records of home site users) whereas the *project supervisors* and the *DEISA supervisor* may only get condensed information.
7. The format of accounting data must comply with standards (OGF UR).
8. A subset of properties that are mandatory for DEISA usage records must be specified (sect. 2.4).
9. Regarding constraint no.8: DEISA partners decided that the product of

the *number of processors* (#CPU) and the *wallclocktime* is the most relevant value for accounting. This accounts also for the usage of idle CPUs that are requested and allocated. The usage of memory, the consumption of storage and of network bandwidth is not (yet) to be considered.

10. The consumption of CPU-time at each machine needs to be normalised by a machine-specific performance index before comparison. Normalised CPU-times can be compared or balanced between machines and sites.

The evaluation of existing accounting systems (see section 1.2) showed that, with regard to the DEISA requirements, significant efforts would have to be invested for integrating any of them smoothly into the DEISA infrastructure. So we developed the accounting facilities based on components that are in use with the DEISA resource monitoring and information system.

2.4 Content of the Accounting Records

The format or the accounting records follow the OGF UR-WG [10] recommendation that specifies 26 base properties of a usage record. Only two properties are mandatory. According to the requirements, the DEISA partners agreed on following minimum subset of mandatory properties (Tab. 1):

Property	Definition
RecordIdentity	unique identifier for each usage record as a compound of the site identifier and a hash value. The hash is calculated from the xml-record excluding the element RecordIdentity
JobIdentity	job identifier assigned by the local scheduler (LocalJobId) and optionally a globally unique GlobalJobId
UserIdentity	Unix username the job has run under (LocalUserId) and the Subject name of the users X.509 certificate (KeyInfo)
JobName	the job name defined by the user.
Status	completion status of the job: completed, failed, aborted, held, queued, started, suspended
WallDuration	the wall clock time that elapsed while the job was running
CpuDuration	total CPU time used, summed up over all processors used
MachineName	name of the compute cluster that executed the job; use in DEISA for assigning corresponding normalisation factors
Host	the name of the host(s) on which the job ran
SubmitHost	the name of the host from which the job was submitted
ProjectName	the name of the project that the job was run under
Processors	the number of requested processors
StartTime	the time at which the usage started
EndTime	the time at which usage ended
SubmitTime	the time when the job was submitted to the system where the job ran. This element has been added to the usage record expecting that this element will be included in a next OGF UR specification.

Table 1: Usage Record properties used by DEISA. Some elements define more than one property

3 Design and Implementation of the Accounting Facilities

3.1 Components of the DEISA Accounting Facilities

The accounting facilities comprise three components (Fig.2): an *accounting data provider*, an *information service* that runs at every DEISA site and a *client tool* for inquiring the *information services*.

Within the framework of the DEISA resource monitoring and information system (RMIS) the so-called *data providers* are running at every site, pushing various resource monitoring data into site-local native XML databases (eXist [11]). These monitoring data are published via *information services* (e.g., Globus MDS). The Java based *data providers* implement the JMX[12] specification (Java MBean technology) that allows to control them remotely.

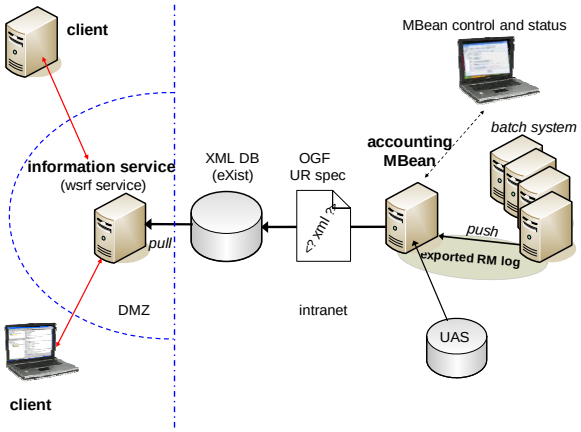

Figure 2: Deployment of the *accounting data provider* (Java MBean) at one site.

The DEISA *accounting data provider* (Fig.2) is based on the same technology. The provider reads the information from proprietary log-files of the resource manager (RM), adds missing property values from the UAS (e.g., project name, X.509 subject name), transforms these combined data into the OGF usage record format [10] and stores it into a collection of the eXist DB. The usage records are periodically pushed into the eXist database, at least once per day.

The availability of accounting information provided in a standard format at every DEISA site makes it possible to uniformly process the data by a generic *report generator*. Such a tool has also been developed and it provides a standardised overview on the monthly consumption of the local compute resources by DEISA users (Tab. 2). The stored accounting records contain CPU times as published by the sites. As different systems have their specific CPU performance these CPU times cannot be compared before they are not normalised with an adequate conversion factor. The normalisation is performed on the client-side based on a centrally accessible document containing these factors. The advantage of this approach is that the conversion can be adjusted without having to change the content of the accounting DBs.

The *accounting information services* are running at every DEISA site allowing authorized entities, such as users, site accounting managers, project supervisors and the DEISA supervisor to retrieve their role-specific subset of

information (Fig.3). The *information services* fetch the accounting data directly from the eXist database on demand, aggregate and filter the information according to the client's role (users and home site accounting administrators may always get the raw data for, e.g., archiving purposes).

```
Site: A  machine: SITE_A_P5                          month 3 / 2007
Project    Nr Jobs    #CPU*WCT[h] wallclockTime[s]    cpuTime[h]
-----------------------------------------------------------------
Project A        2          0.009              34          0.001
Project B        4      35128.178          246995      33159.833
Project C        9          0.022              78          0.003
Project D       10      61440.187         2304010      60647.711
DS-users(csc)   44      53934.337         1518733      47805.130
DS-users(rzg)    2        345.147           38829        343.522
DT-users(bsc)   53          0.168             606          0.021
DT-users(cne)   18          0.039             141          0.006
DT-users(fzj)    1          0.002               6          0.000
DT-users(idr)    9          0.035             125          0.003
DT-users(rzg)    4          0.003               9          0.000
staff(bsc)       1          0.001               5          0.000
staff(cne)       5          5.842            2629          4.495
staff(csc)       2         13.904            1570          6.200
staff(fzj)       9          1.787             838          1.008
staff(hpx)       6          0.179             646          0.130
staff(idr)      31          2.616            2588          1.946
staff(rzg)       4          0.009              34          0.002
-----------------------------------------------------------------
DEISA total    214     150872.466         4117876     141970.013
```

Table 2: A result of the report generator. The product of #CPUs and wallclocktime (WCT) is relevant in order to account also for allocated CPUs that were idling.

Remote parties inquire the accounting data with a *client tool* that opens a SSL connection to the *information services* and authenticates against them via X.509 certificates. Each information service assigns the roles according to the client's identity.

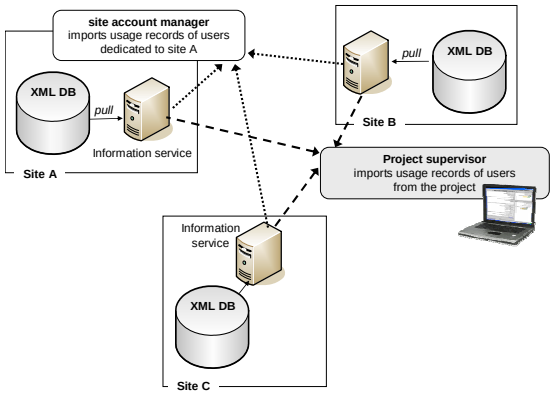

Figure 3: Distributed information services and role-specific views. Exemplary shown here are the *site accounting management view* and the *project supervisor view*

3.2 Accounting Data Provider

The *accounting data provider* runs as a service bean within an MBean container. Fig. 2 depicts the deployment of the *accounting data provider* MBean. The design of the service bean, the *AcctDataPushService* and the associated classes, is shown in Figure 4. The MBean is configurable and the software is designed as an API. It facilitates to develop implementations for specific batch systems as well as for access methods for site-specific local information sources. The abstract class *AcctDataProvider* must be imple-

mented to account for the local batch system. This class provides all the methods needed for creating the proper XML elements according to the OGF UR-WG specification. The XMLBeans framework [13] is used for generating Java classes and methods from the OGF UR schema definition file (XSD). This framework significantly facilitates the adaptation of the software to future revisions of the XML schema definition. Implementations of *AcctDataProvider* have been developed for LoadLeveler and LSF.

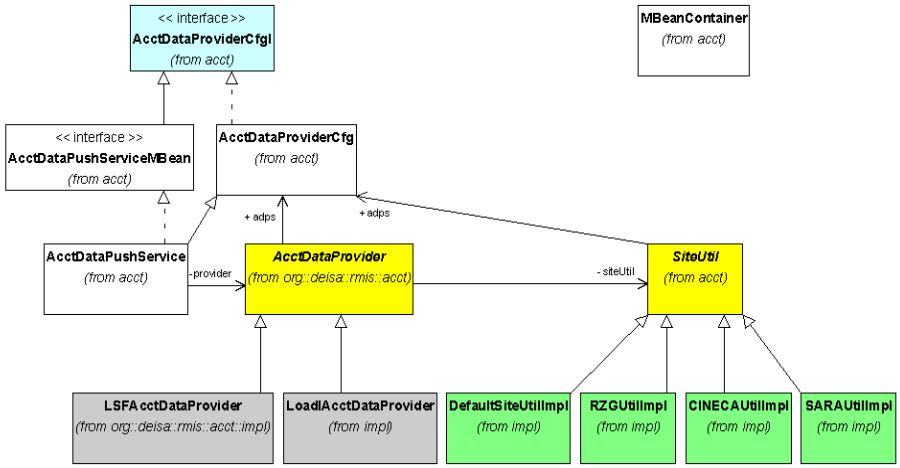

Figure 4: UML class diagram of the accounting data provider software. *AcctDataProvider* and *SiteUtil* are abstract classes. The implementations of the first are specific for each batch system. The implementations of *SiteUtil* provide site-specific methods to access (site-local) sources of additional user information.

Since the scheduler log-files do not always provide all the mandatory UR attributes it can be necessary to retrieve the missing values from other directories or databases. This is supported by the abstract class *SiteUtil* that may be implemented for providing site-specific database access methods. Current implementations (Fig.4) are mainly adding the user's X.509 subject name, project name and the site's machine name to the XML documents.

3.3 Accounting information services and client tools

Currently, a two-track approach is followed to publish the accounting information and to exchange subsets of usage records between DEISA sites. As a temporary solution the data are published via a secured Apache server using a CGI-script that has been derived from the Perl-based report generator mentioned in section 2.1. The accounting data can be retrieved using a GNU Wget-client and reports as shown in Table 2 can also be displayed with a web browser. An X.509 user certificate is mandatory.

More advanced is a WSRF-compliant [14] accounting information service that has been developed using the WSRF core of the Globus toolkit 4 (GTK4). The GTK4 provides certainly one of the most stable WSRF implementations, although the Globus WSRF implementation is based on a late draft of the WSRF specification using an older version of the WS-Addressing specification. A WSRF-client tool for importing the role-specific subsets of usage records has also been developed. It requires only a user's proxy-credential (e.g., created by Globus means). The WSRF-service and client are already in production at some sites for exchanging usage records.

Currently both kinds of information services use a Globus grid-mapfile as an access policy information point. The grid-mapfile is periodically updated by a process that retrieves the mapping information from the DEISA UAS. In addition to the user names for mapping users, the mapping to the roles such as *site accounting administrator* and *project supervisor* are based on a naming convention such as *site-RZG* for the accounting administrator of RZG or *project-CAMP* for a project responsible of the project CAMP.

4 Conclusions

Facilities have been developed for gathering and storing usage records in a common format following the UR-WG specification. Using this format enables the interoperability with other Grid infrastructures. The data is stored locally at each site and the authorised access can also be managed locally.

A strong requirement is the security of the data. On the one hand the data must be well protected due to the legal issues regarding privacy, and on the other hand access must be enabled for those that need the information, at least in a summarized form. The facilities allow for a fine grained management of the authorisation for accessing and viewing the accounting data.

The set up is modular and most facilities can be replaced without disturbing the service, and facilities can easily be adapted according to local needs.

Acknowledgements

We thank all the DEISA partners for their commitment and their contributions, in particular Martin Pels, Luca Clementi, Nicole Lizambert, Vincent Ribaillier, Jarno Laitinen, Anton Frank, Stefanie Meier, and Hermann Lederer. This work has been done under the EU FP6 project no. 508830.

References

1. http://www.deisa.org
2. M. Rambadt et al., "*DEISA and D-Grid: using UNICORE in production Grid infrastructures", GES 2007*
3. http://www.globus.org
4. M.Göhner, M.Waldburger, F.Gubler, G.D.Rodosek, B.Stiller, *Accounting Model for Dynamic Virtual Organizations*, Technical Report, University of Zurich, 2006, ftp://ftp.ifi.unizh.ch/pub/techreports/TR-2006/ifi-2006.11.pdf
5. R. Byrom, et al., *APEL: An implementation of Grid accounting using R-GMA*, http://www.r-gma.org/pub/ah05_364.pdf, 2005
6. The Distributed Grid Accounting System (DGAS), http://www.to.infn.it/grid/accounting/main.html
7. T. Sandholm et al. 2004, *An OGSA-Based Accounting System for Allocation Enforcement across HPC Centers*, http://www.pdc.kth.se/~sandholm/docs/publications/pf081-sandholm.pdf
8. Boston University and NCSA, *Grid-Based Account Management using AMIE*, http://scv.bu.edu/AMIE
9. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri= CELEX:31995L0046:EN:HTML
10. OGF UR-WG, Usage Record Format Recommendation version 1.0, https://forge.gridforum.org/projects/ur-wg, document id 13864
11. http://exist.sourceforge.net
12. Sun Microsystems 2004, http://java.sun.com/j2se/1.5.0/docs/guide/jmx
13. http://xmlbeans.apache.org
14. http:/www.globus.org/wsrf/specs/ws-wsrf.pdf