# The Accounting Infrastructure in EGEE

Pablo Rey[1], Javier Lopez[1], Carlos Fernández[1], Dave Kant[2], John Gordon[2]

[1] Fundación Centro Tecnolóxico de Supercomputación de Galicia (CESGA), Santiago de Compostela, Spain
[2] CCLRC, Rutherford Appleton Laboratory, Oxfordshire, UK

**Abstract.** The EGEE/WLCG CPU accounting infrastructure is based on the collection, processing and presentation of CPU resource usage records that are derived from the log files on the Compute Element (CE) and local batch farm. The data are processed and analysed producing statistical summaries that are available through the EGEE/WLCG Accounting Portal. Access to the data at the Anonymous, User and VOMS level is controlled by means of ACLs. Usage metrics at the VO level are generated allowing for a global overview of CPU resource usage across the entire EGEE/WLCG infrastructure. The analysis of the accounting statistics show that, although there are 118 registered VOs in EGEE, less than 25 VOs are *actively* using the Grid.

## 1 Introduction

The EGEE/WLCG [1, 2] Grid infrastructure has reached more than 30,000 CPU [3] in addition to about 5 Petabytes (5 million Gigabytes) of storage, and maintains 20,000 concurrent jobs on average [4]. There are approximately 3,000 distributed users registered [5] to over 100 Virtual Organisations (VOs). It is clear that this growing infrastructure requires proper accounting in order to know how the resources have been utilised and by whom, and to provide information for effective resource allocation.

The accounting activity collects the accounting data of all sites participating in the EGEE and WLCG infrastructures as well as from sites belonging to other Grid organisations that are collaborating with EGEE. Amongst others, they include the Open Science Grid (OSG) [6], Nordugrid [7], INFN-Grid[8] and GridPP[9].

These data are analysed to generate statistical summaries that are available through the EGEE/WLCG Accounting Portal. These statistics are then available for use, in different views, by: all scientific communities using the named Grid Infrastructures; the owners and administrators of the constituent resources, and the management of the Grid Infrastructures.

The accounting statistics (e.g. CPU job efficiency) available through the Accounting Portal enrich our understanding of the utilisation of grid resources by the different VOs and users/sites. They may be used to identify how resources such as memory, disk and CPU may be better allocated to perform the different

scientific tasks. They can provide input for, and be used to check against *Service Level Agreements* (SLAs).

Storing accounting data from a number of grid infrastructures allows cross-grid projects to see an integrated view of their usage.

The paper is organized as follows: In Section 2 we describe the current EGEE accounting architecture based on APEL and R-GMA. In Section 3 we describe the Accounting Enforcement Activity which aim is to check the integrity of the accounting data. In Section 4 we present the Accounting Portal, the graphical front-end to display the EGEE accounting data. In Section 5 we do an analysis of the accounting data trying to understand how grid is currently being used. Finally, Section 6 we describe the future work.

## 2 Accounting Architecture

In EGEE/WLCG, there are numerous tools available to collect information - both realtime and after-the-event - about resource usage. Widely used but not discussed in this paper, are DGAS [10] (INFN-grid) and Gratia [11] (OSG). The majority of CPU accounting data collected in EGEE is based on the tool APEL which uses R-GMA to transport data across the WAN to a central accounting repository.

### 2.1 APEL

APEL is a log processing application which is used to interpret gatekeeper and batch system logs to produce CPU job accounting records identified with grid identities. It currently supports PBS, LSF, SGE and Condor batch systems and may be extended to support other variants. APEL's implementation is based on a plug-in architecture which separates the core functionality from the actual log parsing (Figure 1).

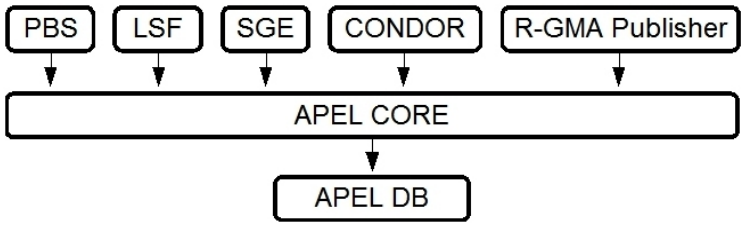


**Fig. 1.** APEL provides a plug-in which parses PBS, LSF, SGE and Condor batch systems logs. A plug-in exists to publish accounting records into R-GMA. Each plug-in connects to the underlying database via the APEL core.

A complete job accounting record is composed of both grid level and local resource information. Amongst other things, they include the submitting users

Distinguished Name (X.509 DN), a unique global job identifier, VOMS group and role attributes, and local resource information (CPU/Wall Clock Time and Real/Virtual Memory) consumed by the job.

This information is typically dispersed between several different log file types such as those produced by the gatekeeper or batch system. For resource usage, a query is issued to the site's BDII (Berkeley Database Information Index) or GRIS (Grid Resource Information Service) to lookup the CPU performance for the computing nodes where the job was executed. APEL attempts to collect all this piecemeal information together and manages it within a *MySQL* database. A further process carried out by APEL then attempts to join the data together to produce a list of final accounting records with all necessary details filled-in. The APEL publisher is used to publish the generated accounting records into R-GMA as shown in Figure 2.

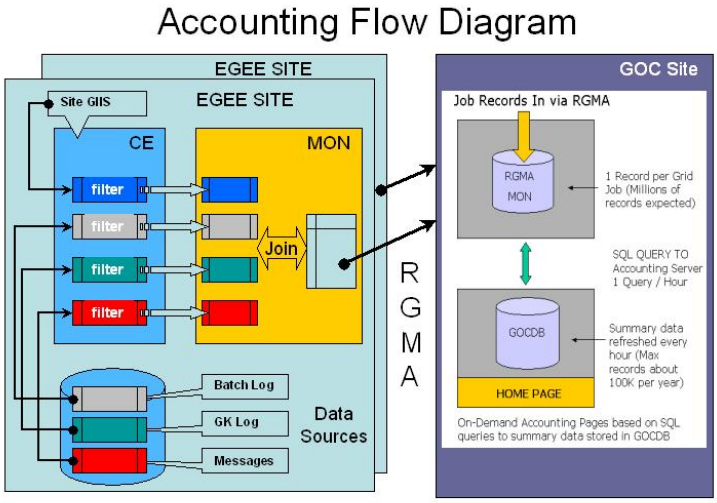


**Fig. 2.** Accounting flow diagram providing a global overview of the data collection process (APEL), and the web reporting service.

APEL provides a mechanism for reliable delivery using a basic integrity check to compute the number of records that were last published compared with the actual count stored on the Grid Operations Centre (GOC). Each accounting record is unique and there is only one record per grid job. The records may be consolidated in different ways to provide high-level views [12] of accounting data. , such as the total CPU time consumed for each EGEE official VO and region, as shown in Figure 3.

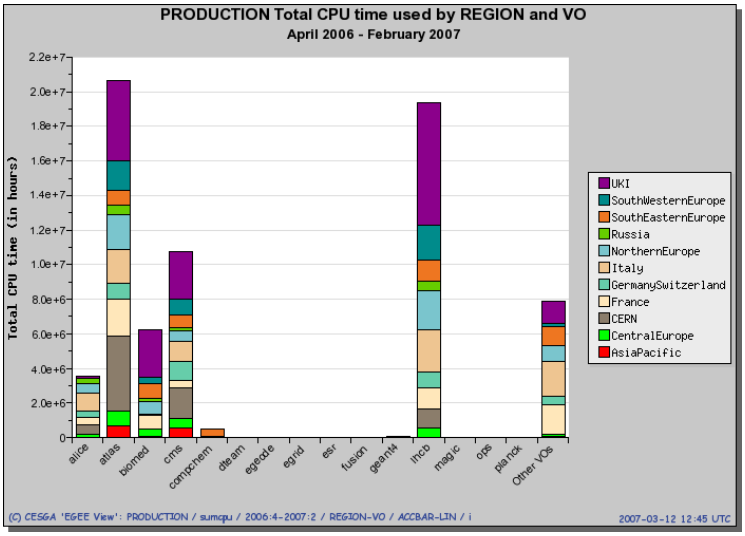


**Fig. 3.** The CPU time grouped by region and EGEE Official VO between April 2006 and February 2007.

### 2.2 R-GMA

The collection of accounting usage records is done through R-GMA [15], an implementation of the Grid Monitoring Architecture (GMA) proposed by the Global Grid Forum (GGF). GMA models the information and monitoring system infrastructure of a grid as a set of Consumers (which request information), Producers (which provide information), a registry, which mediates the communication between the producers and consumers, and a schema which defines the structure of the information to interchange between Consumers and Producers. In EGEE accounting, each site publishes its own accounting data using an instance of an R-GMA primary producer.

To collect the data from all participating sites, data is streamed to a centralised database via a secondary producer that is located at the GOC. Here, the data is processed, aggregated and made accessible to communities via a web portal.

### 2.3 DN Encryption

A breakdown of resource utilisation at the individual user level requires the X.509 DN information to be present in the job accounting record. However, the R-GMA network currently supports Authentication but not Authorisation, so anyone with a certificate issued by an IGTF [13] approved CA can access the data. In order to protect the users identity and maintain confidentiality, APEL performs on-the-fly encryption before streaming the data into R-GMA.

A 1024-bit RSA [14] asymmetric private-public keypair is generated. The public key is used to perform the encryption, while the much larger private key is used to decrypt the data at the GOC. As many jobs can belong to a single user, a randomising function is applied during encryption to reduce the likelihood of repeatable patterns arising in the ciphertex.

The strength of encryption scheme, the size of input data string that can be encrypted, and the speed of the algorithm are important when encrypting data. 1024 bit keys are regarded as strong and a number of cash prizes on offer by RSA laboratories (key challenges) remain open. Unlike encryption with symmetric keys, which can be performed on plaintext strings of an arbitrary size, the maximum number of plaintext bytes that can be encrypted using the public key method depends on the size of the key and the type of padding.

For example, when using a 1024-bit RSA key with PKCS #1 v.1.5 padding (one of the most common options), it is not possible to encrypt a string, which is longer than 117 bytes (that is 117 ASCII or 58 two-byte Unicode characters). Increasing the size of the RSA key to 2048 bits will allow you to encrypt up to 245 bytes of data, but longer RSA key require more time to encrypt and decrypt data. Additionally, the cipher text string generated by a 2048 bit key - after Base64 encoding has been performed - is greater than 300 characters, larger than the character limit allowed in a *MySQL VARCHAR(255)* field.

Thus, a 1024-bit key will not encrypt all DNs. To deal with these rare occurrences (less than 1 % of DNs seen in EGEE), a simple substitution algorithm is applied to shorten the input string before encryption is applied.

In the published accounting data, the longest DNs seen are approximately 117 bytes. Since many jobs can belong the same user DN, successive encryptions may lead to repeatable patterns in the ciphertext. To reduce the likelihood of this, two random numbers n1 and n2 are appended to the X.509 DN before encryption. This increases the length of the input string to 170 bytes.

Following recommendations from RSA laboratories, we use an asymmetric block cipher with padding options: "RSA/ECB/PKCS1Padding" together with the GOC public key. The encryption algorithm also makes use of the SecureRandom function. The output ciphertext is then encoded into Base64 format and can be written to a file or, in the case of APEL, sent as an encrypted stream via R-GMA to the database archiver. The encoded Cipher is 172 characters in length and can be easily accommodated as a database field with limit *VARCHAR(255)*.

```
Input:
/C=UK/O=eScience/OU=QueenMaryLondon/L=Physics/CN=dave kant

Random:
956649486681
/C=UK/O=eScience/OU=QueenMaryLondon/L=Physics/CN=dave kant
30956649486681

Output Cipher text (Base64 String Representation):
Z1R6Lg9CnTr2GZCeJt576Xpe/Dj3EXofwTTGppBzr5nmkt6Vmra7qVqZQXwOcrFi/
```

```
yYE8TYkSPl+RRg4kxdifCBBls1gZf4IB8VKf5Tfa+lOs2cRocd0jgrhOmVpt2h4P
5lGYaJ+zrZYClMmoGjHs2KxOoZH3Tu+cpfUj5XH7RY=

Decryption:
956649486681
/C=UK/O=eScience/OU=QueenMaryLondon/L=Physics/CN=dave kant
30956649486681
```

### 2.4 Intra-VO Level Accounting

A breakdown of resource utilisation at the VO level requires the VOMS Group and Role attributes to be determined for the authorised submitting user by means of the VOMS proxies and the Users Fully Qualified Attribute Name (UserFQAN). The Primary part of the UserFQAN chain is taken to define the VO, Group and Role of the authorized submitting user. The primary component of the chain is used by gLite WMS (Workload Management System) and CE (Compute Element) grid services when resource matching, and it determines the UID and GID that the job is mapped to when it is executed at the local batch resource. On the CE the UserFQAN is obtained from the X.509 DN and the LCMAPS (Local Credential Maping Service) mappings and allows a separation between production managers and ordinary users.

### 2.5 Data Processing

The published encrypted data are removed from the R-GMA network, processed and then inserted into an offline *MySQL* database. The data processing involves the decryption of the X.509 DN, the determination of the VOMS attributes from the primary UserFQAN, and aggregation to form two types of summary accounting data:

- ***Anonymous Statistical Accounting*** This is of the kind "consumed CPU time per VO/ site/ month" and forms the basic building block for rendering accounting data used by the Accounting Portal. This type of data is mostly uncritical and may be provided in a "world readable" way to scientific boards and communities.
- ***User-related*** This type of data can potentially be used to record the work of individual persons. Access to data of this kind must be restricted.

## 3 Accounting Enforcement Activity

The purpose of the Accounting Enforcement Activity [16] is to check the integrity of the accounting data and raise accounting related issues, by means of EGEE GGUS (Global Grid User Support) support and ticketing system [17], to the grid community. The checks performed are:

- ***Site Publishing Data***: A list generated daily that identifies the sites that have not published accounting data in the the last seven days to the GOC.
- ***Identify Orphaned Datasets***: Sites may publish multiple accounting datasets using labels that are not associated to their GOCDB [18] registered sitename. These datasets are grouped together and associated to the GOCDB site. It is important to identify orphaned datasets in order to prevent under accounting of a sites contribution to the VOs.
- ***Meaningful SpecInt Values***: The accounting data may be aggregated together to form usage summaries e.g. total usage per VO per EGEE region. These aggregations involve summations of data across many CPU disparate sites. A simple first order normalisation scheme is applied in order to normalise the data to a common reference scale (1000 *SpecInt2000*). This procedure requires every published accounting record to have an associated *SpecInt* value that is taken from the sites information system. Sites are required to publish a meaningful (non-zero) *SpecInt* value.

### 3.1 Data Synchronisation

Another common problem when accounting for consumed CPU resources is determining whether all the data belonging to a site has been successfully published to the GOC. The APEL publisher performs high-level checksums determined from the local job records at the site and publishes them into R-GMA (*LcgRecordsSync*). They are compared to the corresponding checksums derived from the published records and differences indicate inconsistencies (gaps) in the data record.

## 4 Accounting Portal

The EGEE Accounting Portal [19] is the graphical front-end to EGEE accounting data. Accounting statistics are available through the Accounting Portal for the analysis of the different grid users, VO administrators and site administrators. The Accounting Portal Group (APG) [20] advisory board ensures the development of the Accounting Portal fulfils the requirements of the grid community.

### 4.1 Information Display

The information displayed is of the type *consumed usage per VO/ site/ month* and forms the basic building block for rendering accounting data used by the Accounting Portal. The quantities considered for anonymous statistical reporting are:

- The number of jobs completed.
- The CPU time consumed. This quantity is available in both raw form (as reported by the batch system) and normalised form using a reference scale of 1000 *SpecInt2000*.

– The Wall Clock Time (WCT) elapsed, available in both raw and normalised form.
– CPU Job Efficiency defined as *SUM(CPU) / SUM(WCT)*.

All of the quantities above can be displayed as functions of VO, site or time. In addition aggregation can be done across sites showing (for example) the use by a specific VO as a function of time for all sites in France.

### 4.2 Use Cases

The data building blocks may be consolidated into arbitrary organisational groupings according to different criteria and rendered into trees which are used as a navigation tool for visualisation.

The sites displayed in the EGEE site-groupings are either *Certified* or *Suspended* and, in the Production or Pre-Production (PPS) infrastructure with one or more CEs registered in GOCDB. The *Open Science Grid* (OSG) and any Unregistered site-groupings are also shown in the portal.

**Tier1 View**
A high-level view of LHC CPU resource usage that is comprised of only Tier-1 sites for each of the *LHC VOs* (*alice*, *atlas*, *cms* and *lhcb*). Non-LHC VOs are grouped under a single "*Other VOs*" name.

**Country View**
A geographical view of sites according to their country. This view shows information for each of the *official EGEE VOs* (as defined in the CIC Portal [21]) and the other VOs are grouped under the "*Other VOs*" name.

**EGEE View**
Selecting the Production or the PPS node it is possible to obtain a global accounting view of the project. This view shows information for each of the *official EGEE VOs* and the other VOs are grouped under the "*Other VOs*" name.

It is also possible to obtain a more detailed view of each Regional Operation Center (ROC) or each site. In this case the VOs that have actually run jobs will be shown including *non-official* VOs.

**OSG View**
A View for the accounting information of the sites belonging to the *Open Science Grid* (OSG). In this view the VOs that have actually run jobs will be shown including *non-official* VOs.

**VO Metrics**
The accounting data collected can be used also to generate metrics that allow to assess the status of the EGEE/WLCG infrastructure. The metrics currently available in the Accounting Portal identify aspects related to the usage of infrastructure by the VOs. For example, the *Active* VOs (a VO is considered active if it consumes more than one day of CPU during one week), integrated number of jobs, integrated normalised/raw CPU time and integrated normalised/raw Elapsed time.

**UNREGISTERED View**
Using this view it is possible to obtain the accounting information of the sites that have published accounting data to GOC but are not registered in the GOCDB (*orphaned datasets*). In this view the VOs that have actually run jobs will be shown including *non-official* VOs.

### 4.3 Accounting per User

User DN information is available in the job accounting records but to protect the privacy of the users this information is kept encrypted and the appropriate private key is required to decrypt it (see Section 2.3). Using this information is possible to create statistics about the accounting at the individual user level.

To access to the Accounting per User statistics pages the users need to authenticate using a X.509 certificate signed by an IGTF approved CA. Once the user is successfully authenticated, the authorization could be granted according to the following roles: VO Resource Manager, VO Member, Site Administrator, User and GOC Developer.

Depending on the role assigned the following statistics are accessible:

- **VO Manager View**: Usage for Top 10 Users (Anonomised UserDN), area of pie shows the Total Usage by the VO and the contribution of each of the Top 10 Users and Others, average Wall Clock Time (WCT) for all jobs belonging to each User, ...
- **VO Member View**: The statistics to show in this view are still being discussed with the VOs. This View will allow to have an overview of the status of the VO at a lower level of detail than the VO Manager View (grouping by VOMS *roles* and *groups*).
- **Site Admin View**: Usage for Top 10 Users (Anonomised UserDN), area of pie shows the Total Usage by the SITE and the contribution of each of the Top 10 Users and Others, average Wall Clock Time (WCT) for all jobs belonging to each User, ...
- **User View**: Statistics of usage for all jobs belonging to the UserDN (CPU, WCT, distribution of usage between ROCs and sites, ...)

Initially the plain text UserDNs will not be shown, only the *Anonomised* UserDNs will be shown, due to privacy and data protection issues that might

affect some of countries involved in EGEE/WLCG and that are not well understood yet. The Anonomised UserDN is a modified version of the UserDN that does not allow to identify the user, and which purpose is just to group his/her activity.

## 5 Grid usage in depth

The analysis of accounting data help to better understand how the Grid is being used and to discover different usage patterns between the VOs. This information can help the resource providers and the VO administrators to optimize the utilisation of these resources. For example, to detect overall over-utilisation or under-utilisation and take the appropriate measures.

Studying the trends of a given VO can help to establish those periods with higher activity during the year. During these periods the Resource Centres (RCs) should try to increase the *online* resources to the maximum available. On the other hand, knowing in advance of periods of lower consumption will help the sites to plan their scheduled downtimes, as well as to design alternatives for using the spare resources during these periods.

In this section some examples of such trends are presented, at the same time that we try to answer some basic questions about how the EGEE/LCG grid is currently being used.

### 5.1 How many VOs are using the Grid?

There are 177 VOs that have published data into the accounting database at GOC of these only 118 are registered in the Core Infrastructure Center (CIC) Database. But how many of these VOs are really using the Grid?. Looking at Figure 4 we can see that only a small fraction of these VOs are actually using the Grid, e.g. less than 40 use more than one day of CPU time during one week. Considering a little bit more demanding criterion of consuming one month of CPU time during one week the number of active VOs reduces to less than 25.

An analysis of accounting data published in 2005/2006 by 283 sites and comprised of 25 million job records, show that the usage is dominated by the LHC and biomedical VOs which together consumed a total of 87 % (equivalent to 8269 CPU Years) of the reported usage across EGEE.

### 5.2 How efficient are the jobs running in the Grid?

So it has been shown that only a reduced number of the total VOs are actually using the infrastructure but now the question we want to answer is how efficiently they are using it. For this we define the ratio between the accumulated CPU time and the accumulated elapsed time of the jobs run by a given VO during a given period of time. The results can be seen in Figure 5. It can be seen that only four VOs have a global efficiency higher than 0.75. The two VOs with lower efficiency are *dteam* and *ops* this is reasonable because these are operational VOs used just

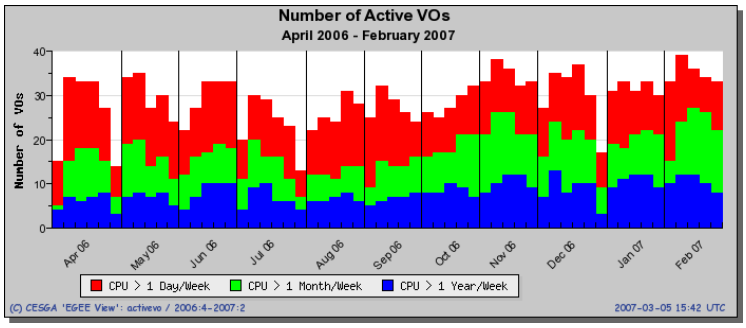


**Fig. 4.** Active VOs: Level of activity per week of each VO of the EGEE project between April 2006 and February 2007.

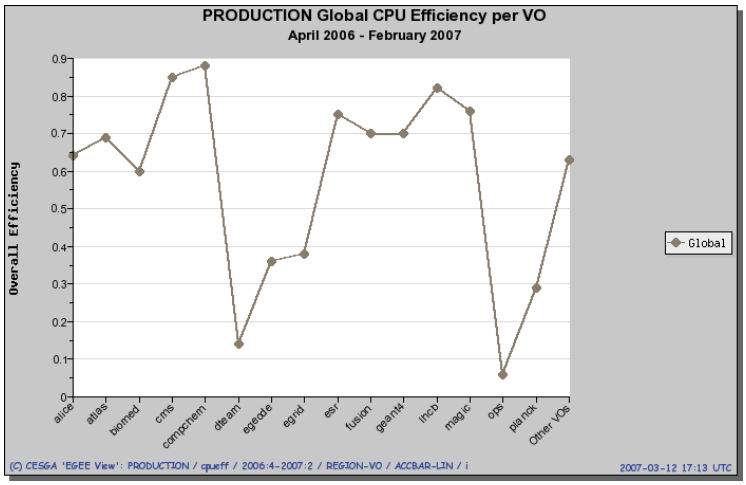


**Fig. 5.** Global CPU Efficiency per VO: ratio between the CPU time and the WCT time of all the jobs run by a given VO between April 2006 and February 2007.

to monitor the infrastructure and not to perform real calculations. Definitively the VOs that show efficiencies below 0.50 should review their jobs looking for possible problems.

### 5.3 Trends of VO Usage

Finally we can analyze how a given VO is using the Grid by looking for trends that help to establish the periods of higher activity during the year. We will focus our attention in the Biomed VO because complete information about their Data Challenges (DC) is available [22].

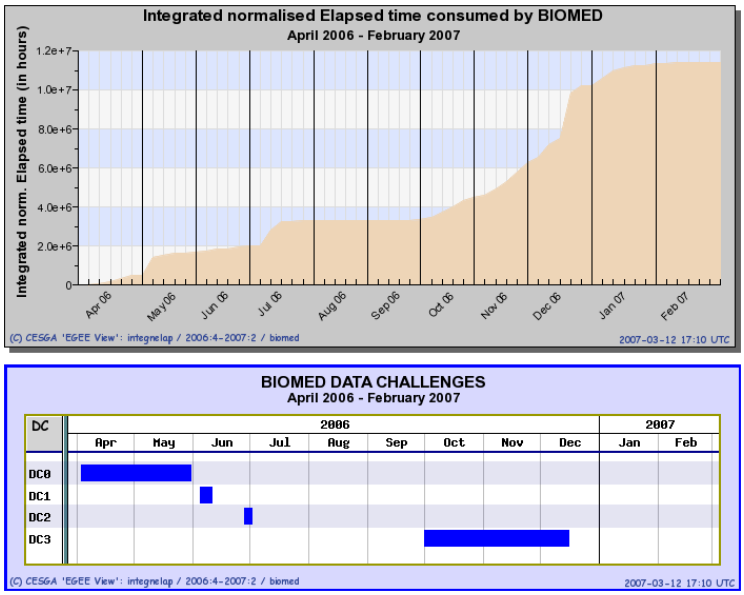


**Fig. 6.** Integrated normalized elapsed time consumed by Biomed VO between April 2006 and February 2007 including Data Challenge information.

Looking at Figure 6 it can be seen that there are certain periods of time when there are steep increases in the accumulated normalized elapsed time consumed by Biomed. This indicates high activity periods where the demand for resources is very high. There are also periods with almost no activity represented by flat regions of the graph. Comparing this graph with the information about the Biomed's DC, shown in the Gantt chart below, it can be seen that the high activity periods match with DC periods. Another interesting fact is that the demand for resources spreads out some time after the official end of the DC. This is an important factor to consider by RCs when planning to contribute their resources to a DC.

## 6 Future work

Apart of the CPU Accounting it could be interesting to get information about the availability and usage of disk and tape storage. A standalone prototype for this has been developed but it needs work to integrate it into a single portal.

The OGF [23] Usage Record is already used but more OGF Standards should be used. Resource Usage Service web services interfaces for data transfer and interrogation are being developed.

Information about the reliability of the infrastructure (cancelled/failed/successful jobs) and the efficiency of the jobs/sites in the Accounting Portal could help to raise underlying problems in the infrastructure or in the resource centers. This information could be include in the portal.

Other useful information that could be include in the portal is about the relation between committed (MoU), installed and used resources. This will help to follow the SLA.

In addition to the accounting data of the EGEE and WLCG infrastructures, the accounting activity is collecting data from other grids that are collaborating with EGEE (OSG, Nordugrid, ...). The Accounting Portal should include the accounting information of any other grid related to EGEE.

This work has concentrated on recording, normalising and presenting the accounting information from the Grid in a common form. No attempt has been made to give any monetary value to the use made of resources. Once the political decision has been made to assign an agreed value to computing, it will be straightforward to record and display this information through our portal.

## References


1. http://www.eu-egee.org/
2. http://lcg.web.cern.ch/LCG/
3. http://goc.grid.sinica.edu.tw/gstat/total/GIISQuery_Usage_cpu_.html
4. http://goc.grid.sinica.edu.tw/gstat/total/GIISQuery_Usage_job_.html
5. EGEE-II QR3: https://edms.cern.ch/document/746593
6. http://www.opensciencegrid.org/
7. http://www.nordugrid.org/
8. http://grid.infn.it/
9. http://www.gridpp.ac.uk/
10. http://www.to.infn.it/grid/accounting/
11. http://gratia-osg.fnal.gov:8880/gratia-reporting/
12. http://goc.grid-support.ac.uk/gridsite/accounting/
13. http://www.gridpma.org/
14. http://www.rsa.com/rsalabs/
15. http://www.r-gma.org/
16. http://www3.egee.cesga.es/acctenfor/
17. https://gus.fzk.de/
18. https://goc.grid-support.ac.uk/gridsite/gocdb2/
19. http://www3.egee.cesga.es/gridsite/accounting/CESGA/egee_view.php
20. http://egee-docs.web.cern.ch/egee-docs/list.php?dir=./apg/&

21. https://cic.gridops.org/index.php?section=vo&page=homepage
22. https://cic.gridops.org/index.php?section=vo&page=datachallenges
23. http://www.ogf.org/
24. EGEE Glossary: http://www.eu-egee.org/introduction/EGEEGLOSSARY