Notes from the Workshop on new types of Accounting at EGI Technical Forum 2011

Lyon, 22nd September 2011
John Gordon

This workshop aimed to consider the requirements and plans for new types of accounting in EGI. While this is part of the workplan of EGI-Inspire JRA1 for years 2-4, there are obviously many other stakeholders and this workshop aimed to bring many of them together. The agenda and slides are available here http://www.egi.eu/indico/sessionDisplay.py?sessionId=89&confId=452#20110922

The types of accounting considered, and the stakeholders were:

- CPU (inc OGF UR) (EGI, EMI, OGF)
- MPI (EGI, EMI)
- Storage (inc StAR) (EMI, OGF, EGI)
- Virtualisation (EGI, other projects)
- Applications (EGI)
- Data Use (EU-DAT, PaNData, ...)

These notes will not repeat the content of the various talk given but just highlight issues and conclusions.

CPU

EMI is reviewing the OGF-UR with the aim of harmonising use within EMI products. All current systems are based on OGF URv1 but differ in semantics and extensions so usage records are not interchangeable. APEL is participating and will track its output. Others should be informed through the OGF UR-WG and other routes (Action) David Wallom claimed there are now 8 implementations of the OGF UR. These should be traced (Action)

Current APEL schema receives raw cpu with a benchmark value which is the average published for the cluster. This is not exact. If sites are charging money for cpu use then the records have to be exact. Proper use of the GLUE subcluster could solve this (FNAL already use this). The glite-cluster should solve this for CREAM. Check how Unicore and ARC handle this issue (Action). Some batch systems record the power of the node(s) used by a job in their batch logs. This would allow a client to publish the exact benchmark for each job.

MPI

Input from Enol Fernandez and others. OGF UR already contains fields for the number and names of nodes used by a job and the number of cpus. These are not currently implemented in APEL but are in the revisions currently being implemented (Action). It is believed that relevant batch systems provide the information required - names of hosts and cores used.

Since there is use of OpenMP as well as MPI, it should be all parallel jobs that are accounted, not just MPI.

No extensions should be required to the UR for parallel jobs but consistency in recording the number of cores used in the field for the number of cpus was agreed. There are two reasons for this: (a) consistency with benchmark/core and core-hour used for single-threaded jobs; (b) to record multiple threaded jobs even on a single cpu or node.

The portal needs feedback on how parallel jobs should be displayed there. Currently individual jobs are not exposed, just summaries for site/VO/user/month. The average number of cores used in a job would not be very meaningful if it included ncores=1. Perhaps separate presentation of jobs with nodes or cores >1 would be useful.

Storage

Presented by Jon Kerr Nilsen. The Storage Accounting Record (StAR) was discussed again at OGF earlier in the TF. Plan to complete a document for public comment by the end of the year which would allow agreement of it at OGF 34 in Oxford in March 2012.

JG was concerned that the current record only recorded disk space occupied by files and not allocated, or the total size of the space defined for the user/VO. This is equivalent to accounting only for cpu usage and ignoring wallclock. There are use cases for both. He would pursue this issue (Action)

It is the intention of the EMI storage products to implement accounting in their EMI-3 release in May 2013. JG will attend their AHM in October to present the EGI APEL infrastructure as a means of gathering these records (Action). Other, non-EMI storage products should be contacted (Action) as storage accounting will be of little use unless all storage on the Grid is accounted.

Virtualisation

Presentation by David Wallom on the accounting work in FLESSR, a UK project on federated clouds. They collect far more information than is required in a usage record so a strawman UR will be proposed (Action) so that (a) the feasibility of producing this UR from all VMMs can be evaluated; (b) users and other cloud projects can comment on whether it meets their needs and expectations.

A conclusion of the Cloud and Virtualisation Workshop held by EGI in June 2011 was that accounting of a VM instance should not drill down into the use made of the VM by individual users but restrict itself to the total usage by the VM. Thus an instantiation of a VM has many similarities with a job and the existing OGF UR for jobs would be a good starting point. The use case of a long running VM being accounted regularly before it ends is already met by the existing UR which has a status field as it was always anticipated that intermediate usage records could be cut during execution of a job. Most existing accounting systems haven't implemented it though.

Applications

Ivan Diaz presented on CESGA work. No concrete use cases but believe that memory usage is important. Need a method of recording which application is being recorded. Binary or script name is probably not enough. One possibility is to use VOMS groups for applications. This would require coordination across VOs to be useful but is worth further study.

Gratia from FNAL, used by OSG, has process level accounting at the site level which can be interpreted to reveal which applications are being run. Unicore also record the process but this reveals a lot of 'bash'. Hannover have deployed a local system which is worthy of more investigation.

Data Use

Storage accounting will record the ownership of data but not who uses it. Large data centres wish data on use of their resources either for charging or for justification of their existence, the relative usage by their user communities and for capacity planning.