Grid Accounting and Auditing
Joint Project between Fermilab, PPDG, SLAC, US CMS,
Contributed as an OSG Activity
Project Definition

Authors: Sudhir Borra, Philippe Canal, Matteo Melani, Ruth Pordes
Version 1.3

Short Description

- The Grid Accounting and Auditing Join Project (contributed as an OSG Activity) designs and deploys robust, scalable, trustable and dependable grid accounting and auditing services, publishes an interface to the services and provides a reference implementation.

Goals

The main goal of the Grid Accounting and Auditing Joint Project, which is additionally being contributed as an OSG Activity, is to provide the stakeholders with a reliable and accurate set of views of the Grid resources usage.

The Grid Accounting Project will:

- design the schema for the accounting attributes,
- ensure the necessary collectors and sensors are in place in the resource providers,
- define and deploy repository and access tools for the reporting and analysis of the grid wide accounting information.

The Accounting system will properly determine a confidence level in the existing accounting information and adequately address and present erroneous or missing accounting data.

The accounting system will adequately protect the privacy of the users and organizations involved.

The auditing system will use information from the accounting system and link it to information from other sources to allow full tracking and analysis of the actions and events related to a user’s resource usage.

The auditing system needs to be able to present the immediate and short term information of the state and transitions in a user's use of a resource.

The initial main goal for the accounting system will be to track VO members' resource usage and to present that information in a consistent Grid-wide view, focusing in particular on CPU and Disk Storage utilization.

References
This document uses material from OSG Accounting Requirements
A preliminary survey of tools in use on Grids for accounting has been done, but not included in this document: R-GMA, MonaLisa, APEL

General Description

The Accounting system can be decomposed in 4 separate but related parts.

- Collection of the accounting data
    - Currently partially done by tools like Ganglia and MIS-CI and by parsing the content of other tools' log files
    - One of the responsibility of the Accounting and Auditing OSG Activity will be to help shape and define the data being collected
- Interface between collectors and the accounting 'database'
    - This will be the set of well defined and simple to use interfaces used by the collectors and other process, storage or networking management tools
- Set(s) of distributed data stores containing the accounting information
    - The resulting database needs to be able to support the physical segregation of the accounting data per facility, per VO and per GRIDs. It also needs to support the proper authorization checks.
- Interface to publish the accounting information contained in the databases.
    - This is an interface that allows to aggregate accounting information from different resource providers, and therefore to create centralized databases of accounting records.
- A presentation layer giving access to report and aggregates information

  - This needs to support the proper authorization checks before providing the information. For example, a regular user should be able to access the accounting information about her jobs but nobody else in particular.

The Auditing system will make use of the accounting infrastructure and link to additional data to enable full tracking and traceability of a users actions and resource use.

Stakeholder and Users

We base our definition of following terms on the Open Science Grid definitions:

VO (the Globus definition):

"Virtual Organization– A dynamic collection of Users, Resources and Services for sharing of Resources (Globus definition). A VO is party to contracts between Resource Providers & VOs which govern resource usage and policies. A subVO is a sub-set of the Users and Services within a VO which operates under the contracts of the parent"

- End User:

  A person who makes a request of the Open Science Grid infrastructure.
- Organization:

  A group of people from a subset of brick & mortar institutions (not necessarily a contiguous subset)

We expect the following type of usage:

- End users.
  - Accounting: Would like information about how much GRID resources they consumed, in particular in relation to they allotment.
  - Auditing: Would like information as to the state and state transitions associated with their resource use
- Facility managers
  - Accounting: Would like to know who is using their resources and how much.
  - Accounting and Auditing: Would like to be able to know accurately what the current utilization of the resources is and what the trends are. They can use this information to better plan the future of their facility.
  - Auditing: Would like to trace the state and state transitions of a user’s use of a resource in order to understand both normal and anomalous behaviors.

- VO managers
    - Same as facility managers but for resources they own or provide
    - Would like to know how their members are utilizing the GRID resources provided to them.
- GRID managers
    - Same as facility managers but for resources they own, provide.or oversee.
- Organizations
    - Same as facility managers but for resources they own or provide.
- Resource Providers
    - Same as facility managers but for resources they own or provide.
- Auditing Agent
    - Wants to perform an audit to verify the accounting information.
    - Wants to investigate an incident in more details.

The direct stakeholders of this project are: US CMS, Fermilab and SLAC Computing Divisions, and the PPDG Common Project. Through the project contributions to the Open Science Grid US ATLAS and OSG are also stakeholders.

Workflow

Stakeholders will be able to generate accounting reports using the Grid Accounting system. The Grid Accounting system will also offers an interface through which stakeholders will be able to export accounting data for custom manipulation and presentation. The project will work closely with the Stakeholders to develop, implement these reports. The project will be responsible for providing the interfaces needed and initially for posting and archiving these accounting reports to a web accessible repository. Stakeholders will be able to use this web repository to review and transmit the reports as needed.

Current Status

PPDG effort in the accounting project was started at the beginning of 2005 to extend the existing MonaLisa monitoring system to monitor storage space at OSG sites. A new MonaLisa module called VoStorage has been created to monitor the utilization of the $APP, $ DATA and $TMP directories. The VoStorage reports disk utilization per VOs.

An accounting system prototype is under development: The Gaspacio Server. This is a stand alone server written in Java that accept accounting records in XML formats, cache them in memory and allows accounting data consumers to retrieve the accounting data through in XML format. The goals of the prototype are to better understand the accounting problem and to get familiar with the Java, XML and Web Services technology.

In OSG 0.2 accounting is based on the information in the MonaLisa repository. In Grid3 this information was analyzed to give MDViewer plots. These plots were based on integration and analysis of the MonaLisa data. Exactly what MonaLisa plots will be used for OSG accounting in the next few months is still being debated.

The ACDC Grid dashboard gives additional information about the status and performance of the OSG.

Effort
The current effort available to this project is:

Philippe Canal	Fermilab CD	.5 FTE	~.1FTE til ROOT backfill available
Sudhir Borra	US CMS	.75 FTE	>=.25FTE for current OSG support
Matteo Melani	SLAC, PPDG	.5 FTE	Multiplexed with support for OSG site at SLAC.

Timelines
Initially, the prototype accounting system will focus on providing more accurate CPU and storage utilization reports.

- Spring 2005: research and analyze the existing solution
- August 2005: First release of the Accounting Interfaces, including adaptation of the Monalisa VO callout. These interfaces are the part of the system directly seen by 'director processes' (data providers, sensors, jobs and storage managers, etc.). Hopefully, the interface will be ready for the director processes developers to start developing using the Accounting Interfaces. The Accounting Interfaces will be flexible and extendable enough to support the necessary extension of the schema and properly hides the implementation details of the Accounting Data stores. At the same time a prototype database and presentation layer will be released, both will be based on existing tools and databases and are not expected to meet the full requirements of distributiveness or privacy.
- December 2005 or January 2006: Second major release of the Accounting System, this should include an improved database (especially if we can actually leverage existing infrastructure) and proper hooks for authentication to access the data.

- In parallel to these developments we plan on focusing starting June 2005 on the extension of the schema and the analysis needed to properly understand and deal with missing or erroneous data. This part will heavily really on the support and help of the developer of collectors, sensors and director Grid processes.

Collected Data and Database Schema

The initial basis of the accounting database schema will be provided by the Usage Record Working Group from the Global Grid Forum and inputs from the OSG stakeholders.

This schema also needs to be rich enough to provide the information needed to effectively collaborate and interface with other peer GRID (LCG in particular).

General notes

Currently the collection of accounting information is mostly based on collectors processes which grab their information from log files produced by the local workload and storage managers. These collectors have an inherent dependency on the format of those log files. These collectors have been historically developed by groups other than the maintainers of 'the local workload and storage managers'.

This technique is intrinsically brittle given the fact that the format of the log file is often not considered essential to the developers and is subject to unannounced and/or undocumented changes.

If the accounting information is considered to be central to the operation of the GRID(s), it is essential that we shift the paradigm and have 'the local workload and storage managers' directly feed the information into the 'accounting system' rather than having to collect it from the log files.

As a general rule, the Accounting and Auditing OSG Activity should strive to reuse as much as practical the existing infrastructure and software.

Operations of the Accounting System

This document does not address or describe the work needed to operate the Accounting system. We will respond to requirements from the different users and managers to implement a Business Process to report the information that they need. We need to understand these well enough so that the Accounting system is rich enough to support them properly.

Interfacing to the Existing Laboratory Accounting Systems

The Accounting system will interface with, rely on or correlate with the existing accounting systems.

Interfacing to the EGEE and LCG Accounting

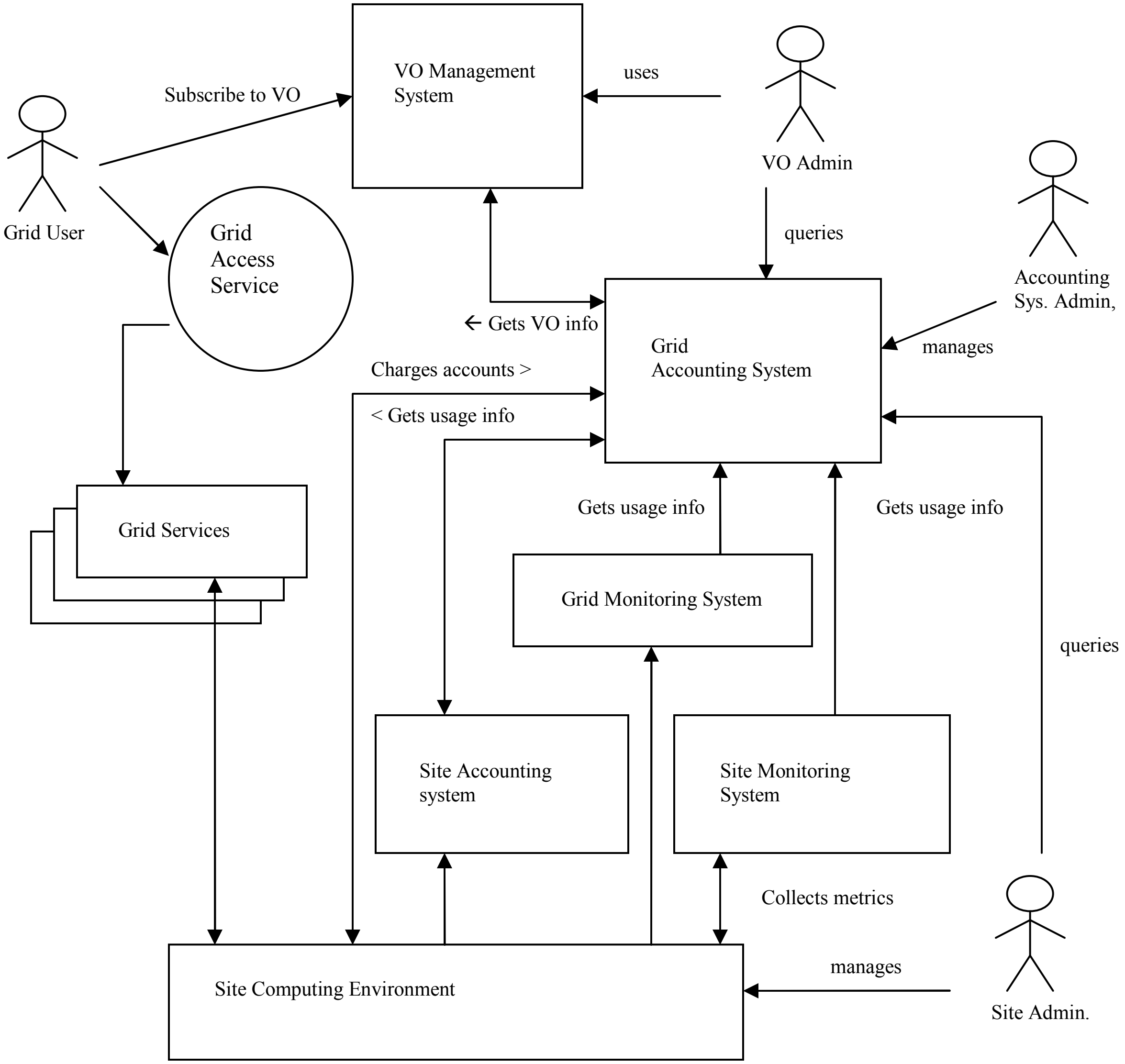
The EGEE accounting in based on a modified schema derived from the GGF Usage record working group. The information is presented to the GOC Accounting Service:

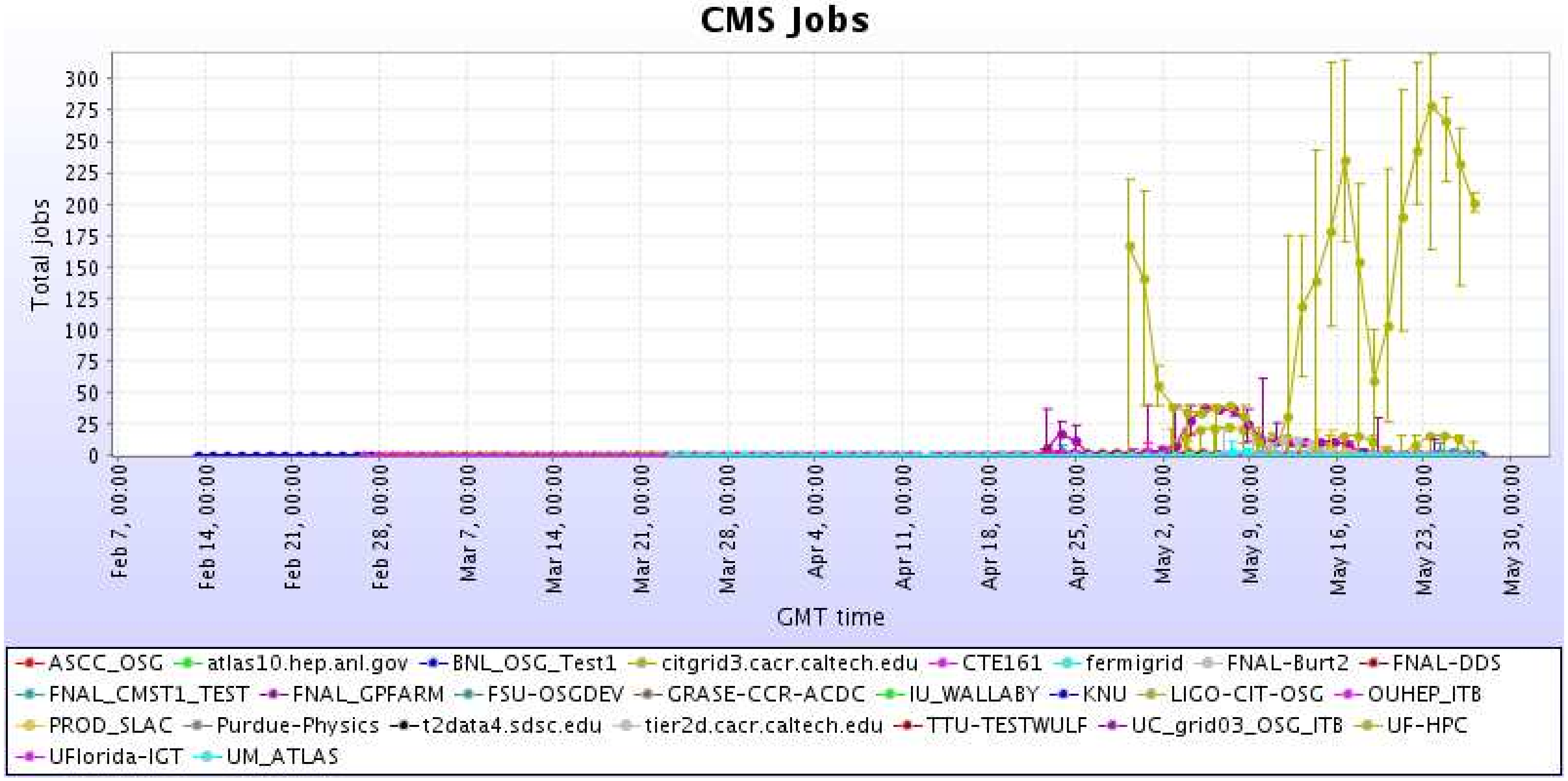
- Sites install an RGMA MON node and deploy the accounting program onto the CE.
- Program parses log files, writes accounting data to the MySQL database on the MON node.
- This data is streamed to the GOC RGMA Accounting server
- The GOC accounting database contains one row of data for every successful job sent through a PBS (or BQS) job manager at each site.
- The data is consolidated (per site, per VO, per month) and presented at the web site.

Accounting is supported for the PBS and LSF batch systems.
Accounting information can be transmitted to the LCG in the following format:

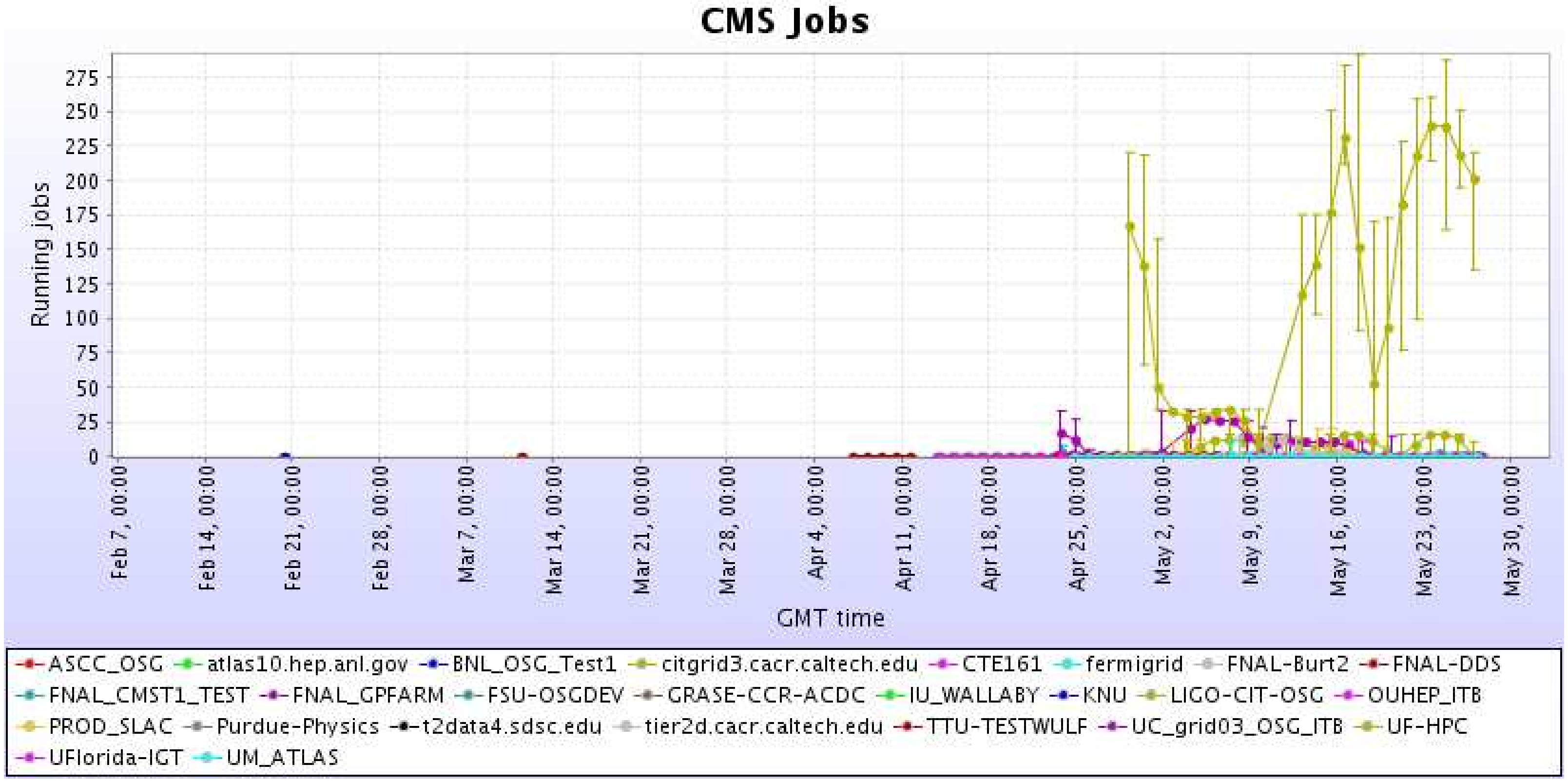
Field	*Type*	*Null*	*Key*	*Default*	*Extra*
ExecutingSite	varchar(50)		PRI		
LCGUserVO	varchar(50)		PRI		
Njobs	int(11)	YES		NULL	
SumCPU	int(11)	YES		NULL	
NormSumCPU	int(11)	YES		NULL	
Month	int(11)		PRI	0	
Year	int(11)		PRI	0	
RecordStart	date	YES		NULL	
RecordEnd	date	YES		NULL	

SumCPU represent unnormalised CPU time in units of hours.
NormSumCPU represents normalised CPU time in units of hours. We normalise to 1K.SI2K across LCG.

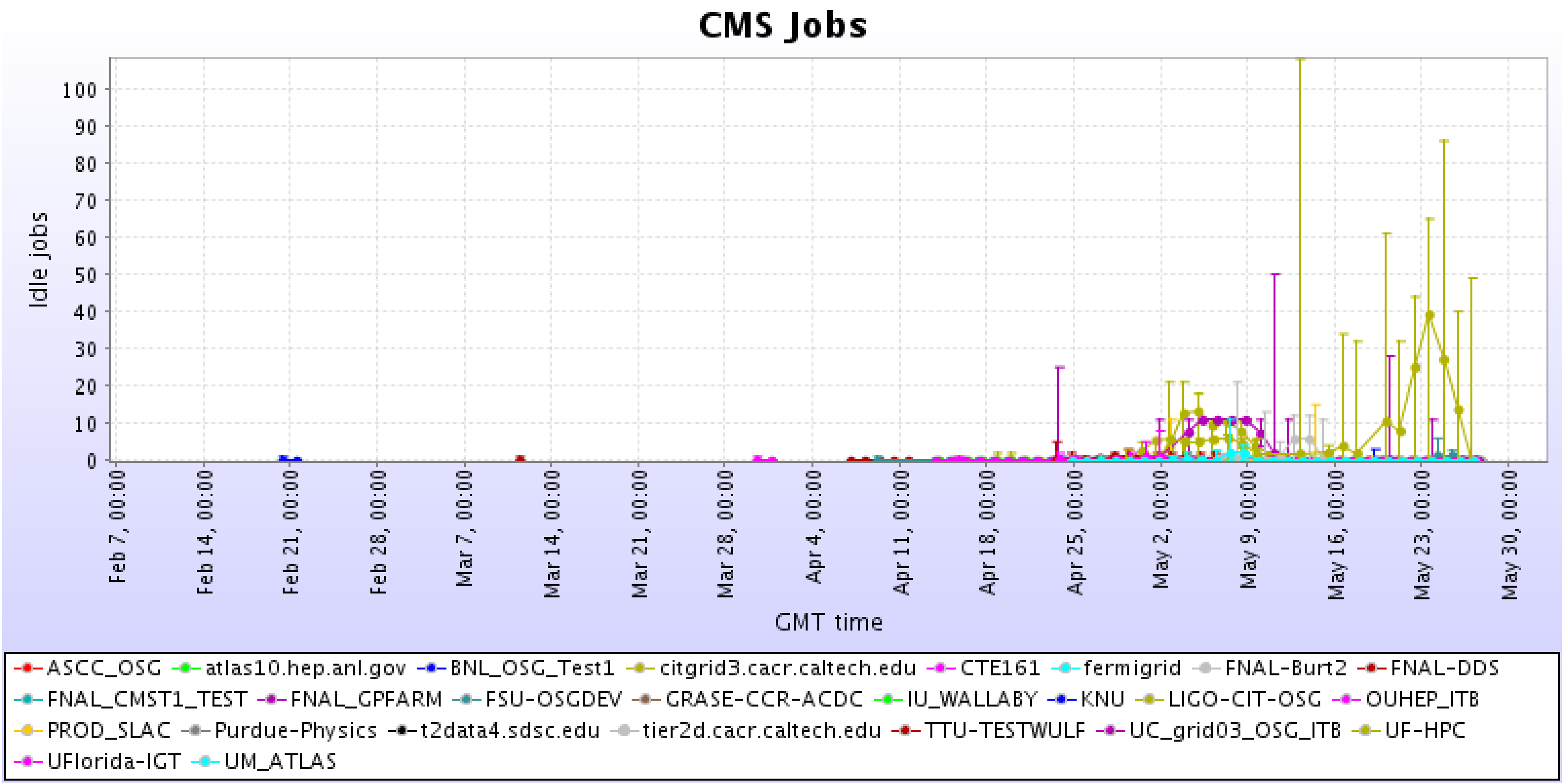

VO Management System
uses
VO Admin
Subscribe to VO
Grid User
Grid Access Service
queries
Accounting Sys. Admin,
← Gets VO info
Grid Accounting System
manages
Charges accounts >
< Gets usage info
Grid Services
Gets usage info
Gets usage info
Grid Monitoring System
queries
Site Accounting system
Site Monitoring System
Collects metrics
manages
Site Computing Environment
Site Admin.



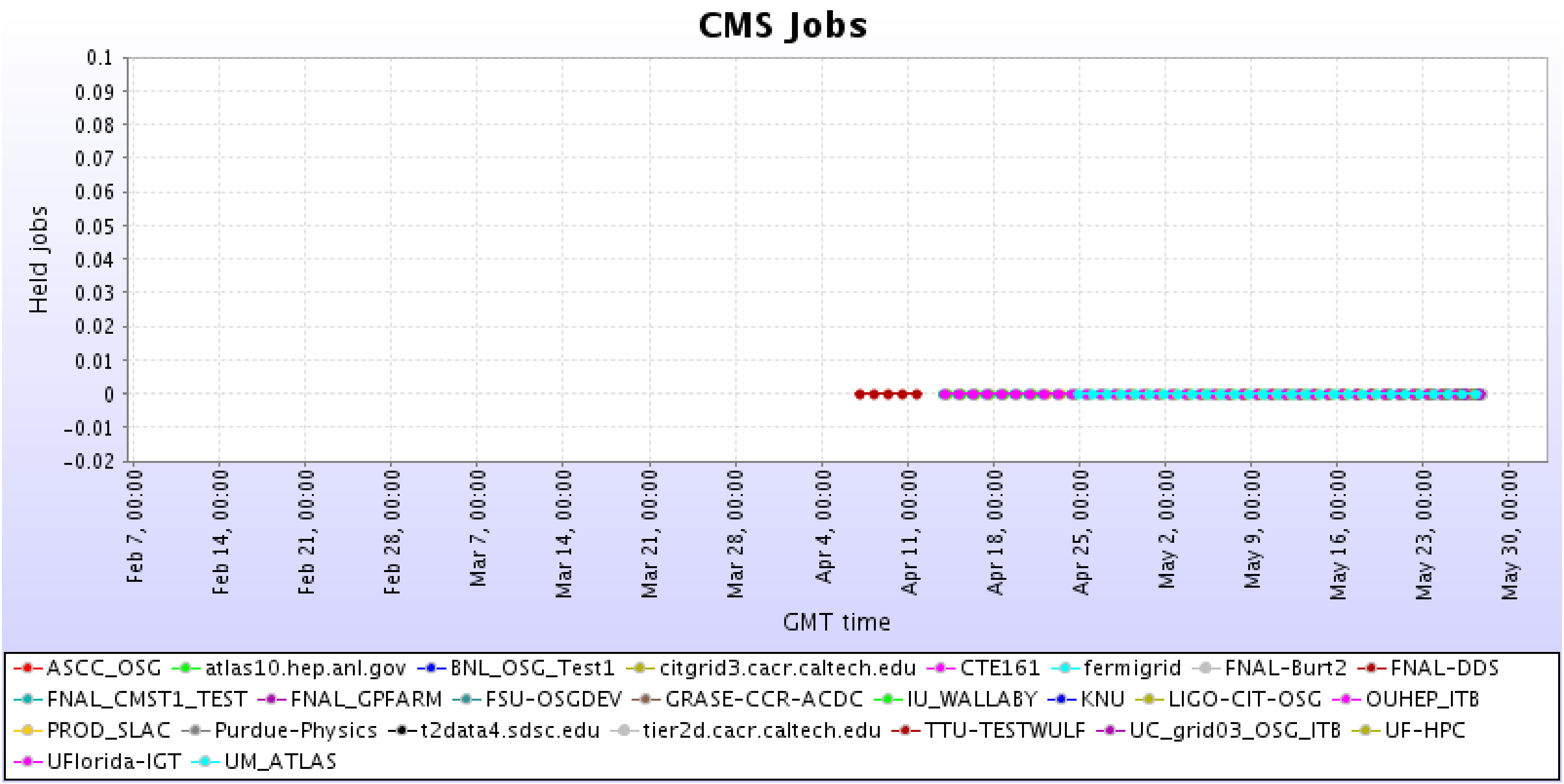
CMS Jobs Submitted over a period of time (4 months in the above graph) .Note that the graph is a “simple” graph and not cumulative.



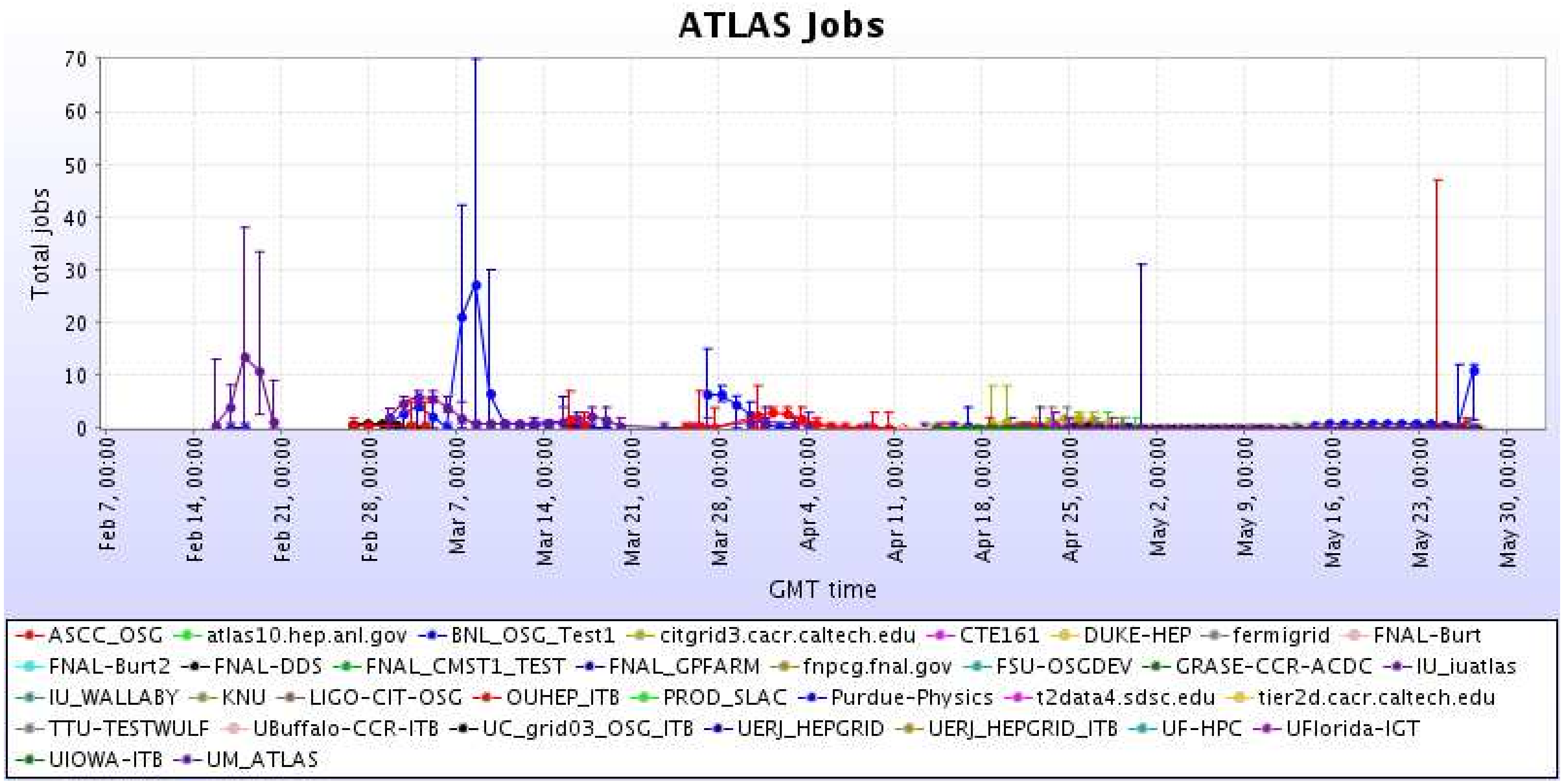
CMS"Running Jobs" over a period of time (4 months in the above graph) .Note that the graph is a "simple" graph and not cumulative.



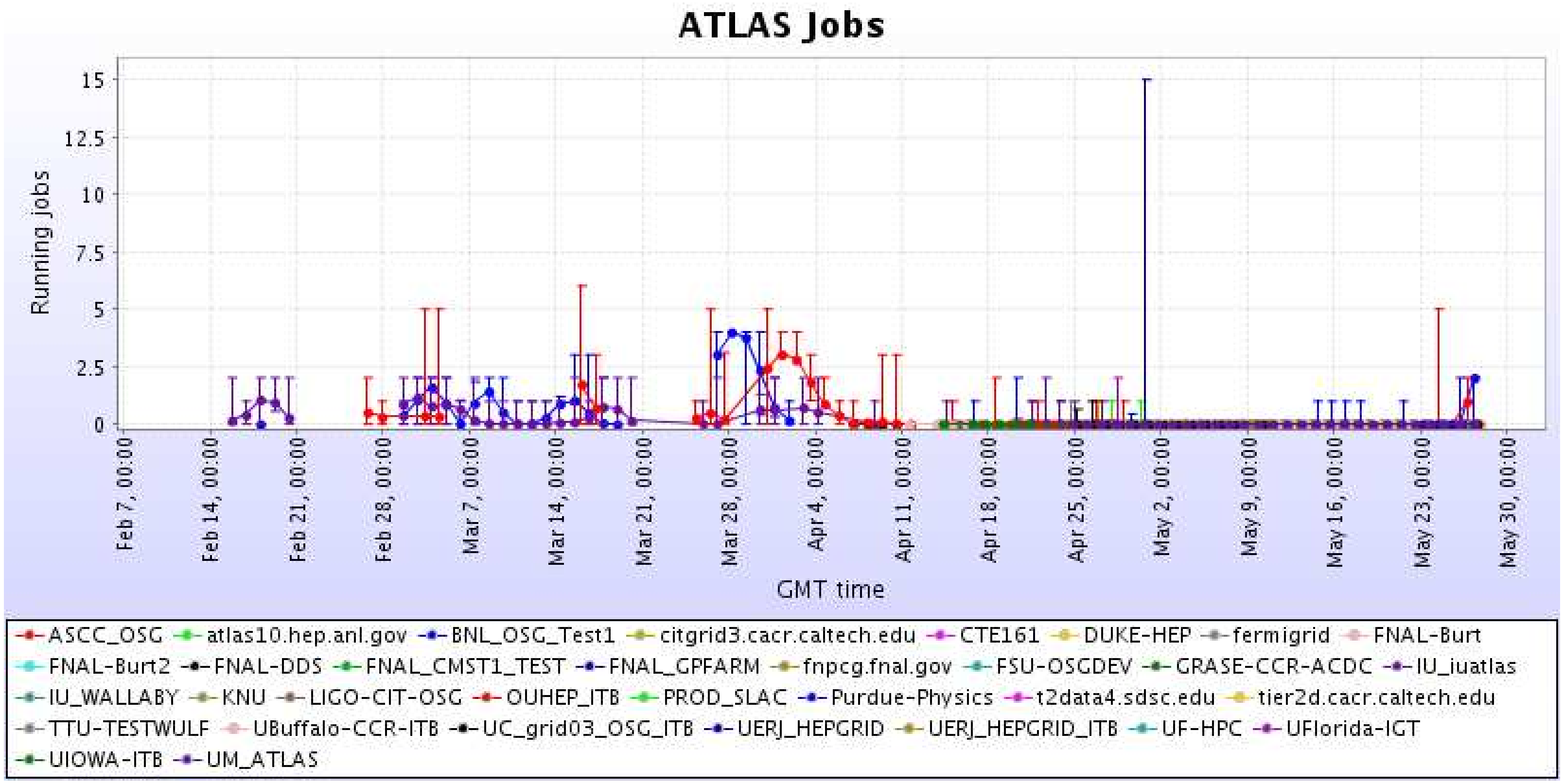
CMS ”Idle Jobs” over a period of time (4 months in the above graph) .Note that the graph is a “simple” graph and not cumulative.



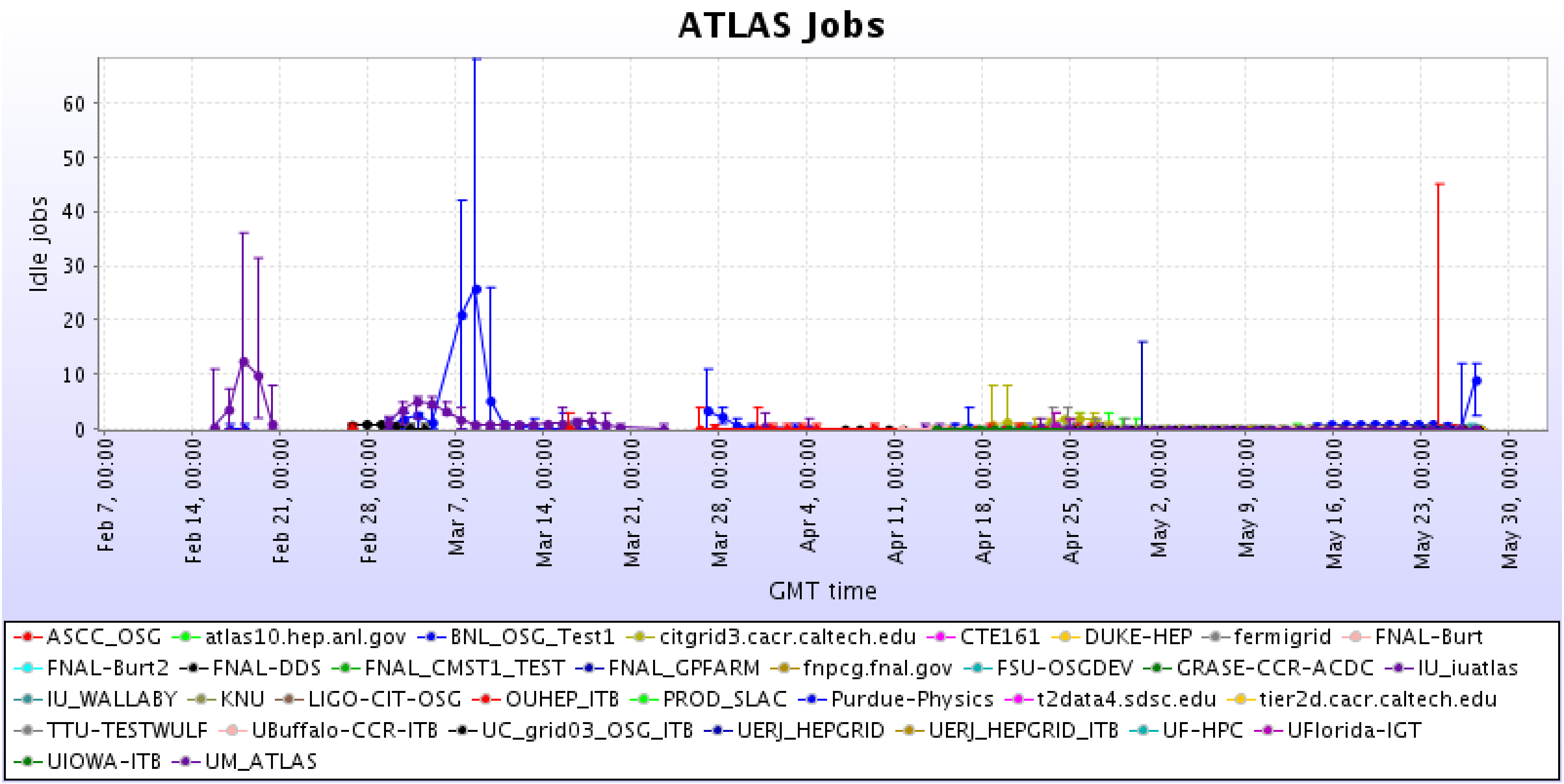
CMS ”Held Jobs” over a period of time (4 months in the above graph) .Note that the graph is a “simple” graph and not cumulative.



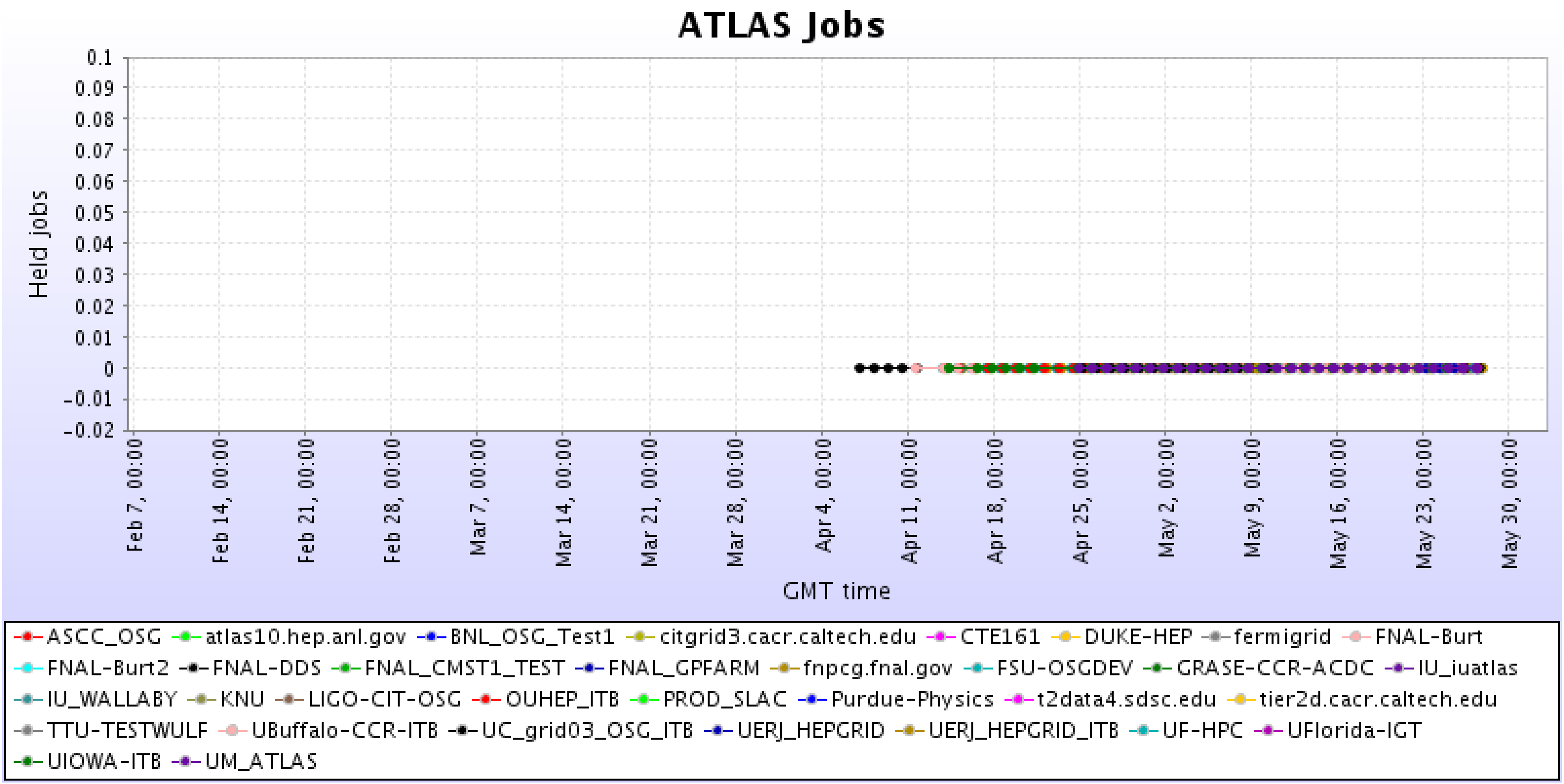
ATLAS Jobs Submitted over a period of time (4 months in the above graph) .Note that the graph is a “simple” graph and not cumulative.



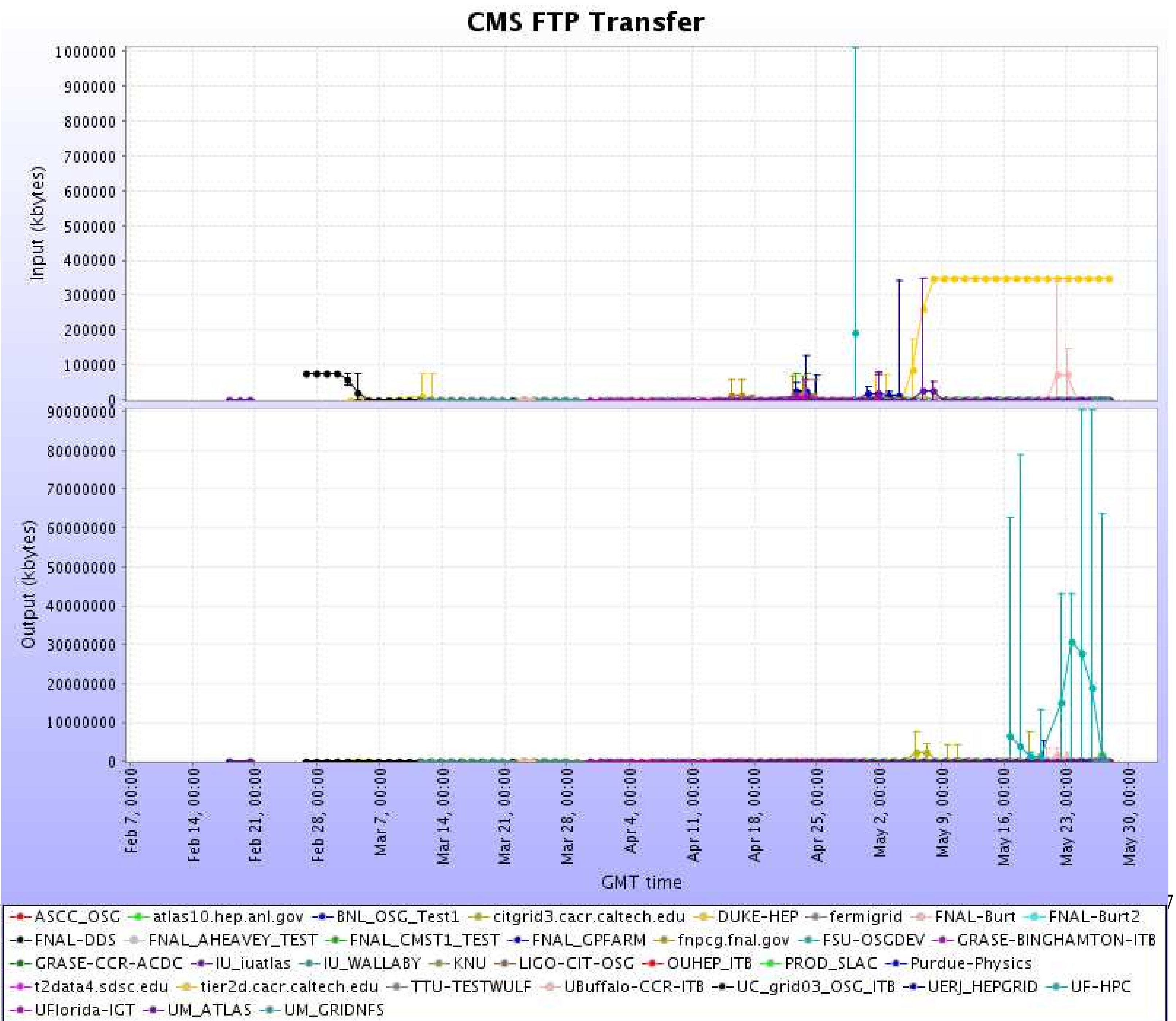
ATLAS ”Running Jobs” over a period of time (4 months in the above graph) .Note that the graph is a “simple” graph and not cumulative.



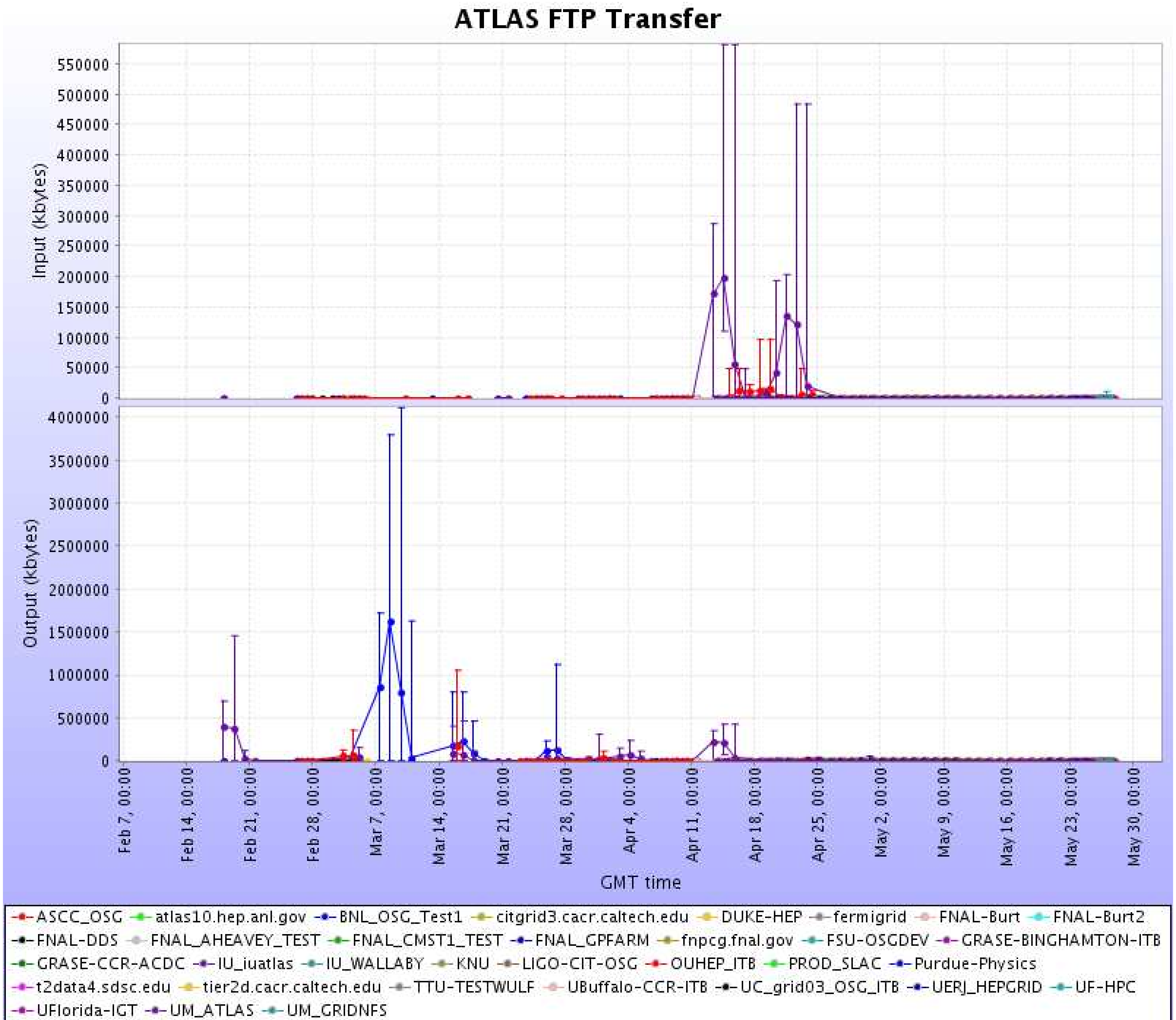
ATLAS ”Idle Jobs” over a period of time (4 months in the above graph) .Note that the graph is a “simple” graph and not cumulative.

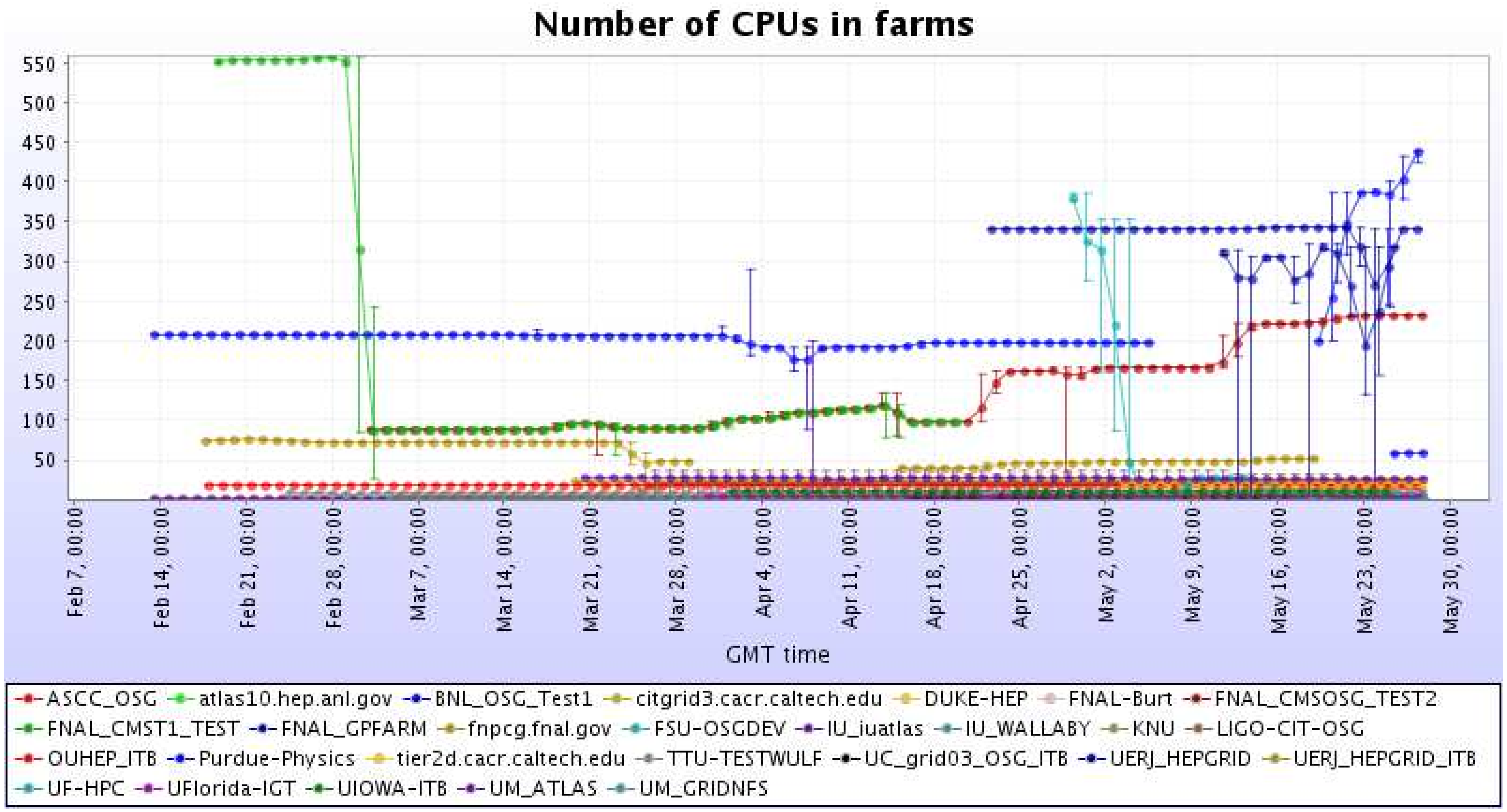


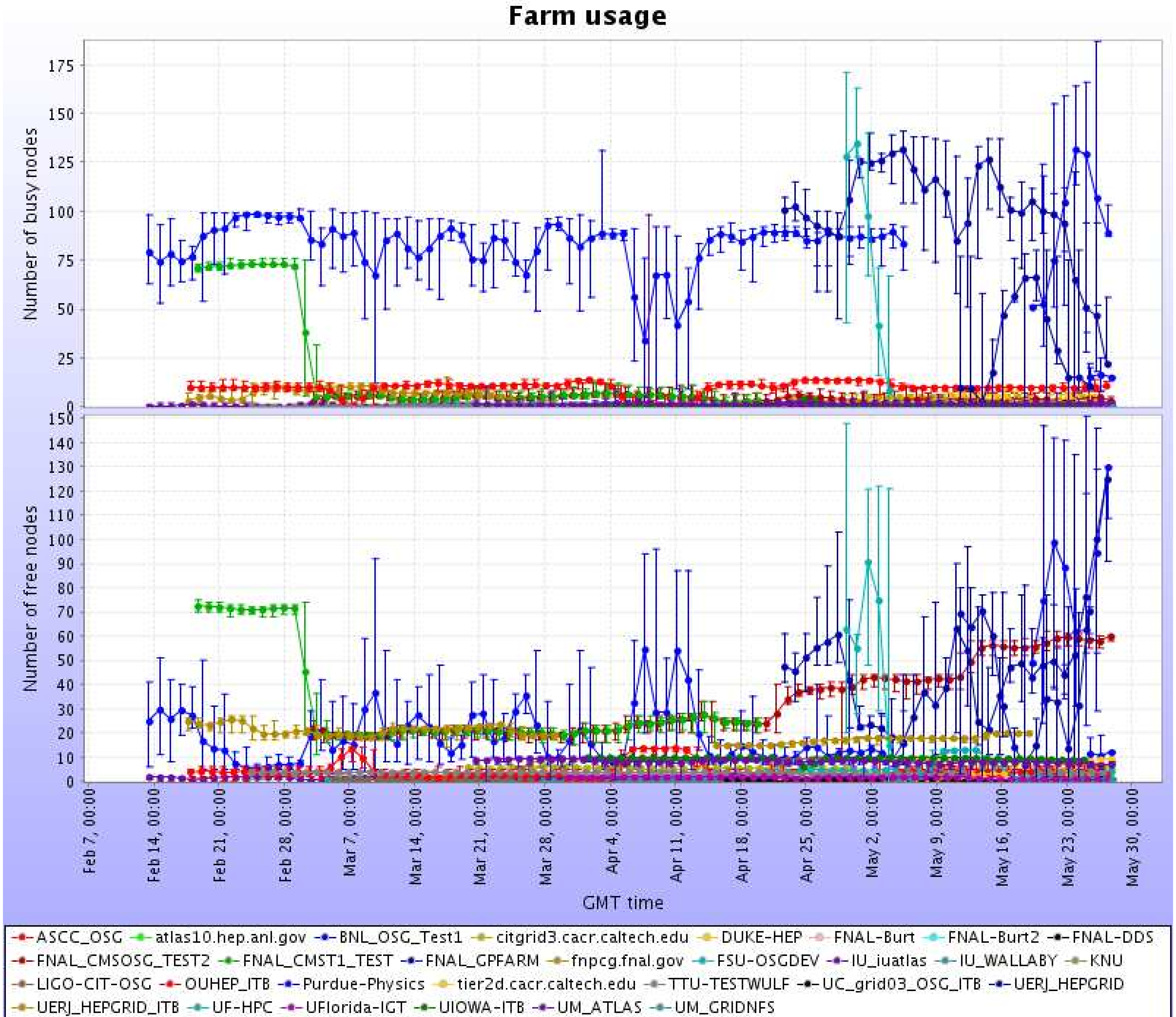
ATLAS ”Held Jobs” over a period of time (4 months in the above graph) .Note that the graph is a “simple” graph and not cumulative.

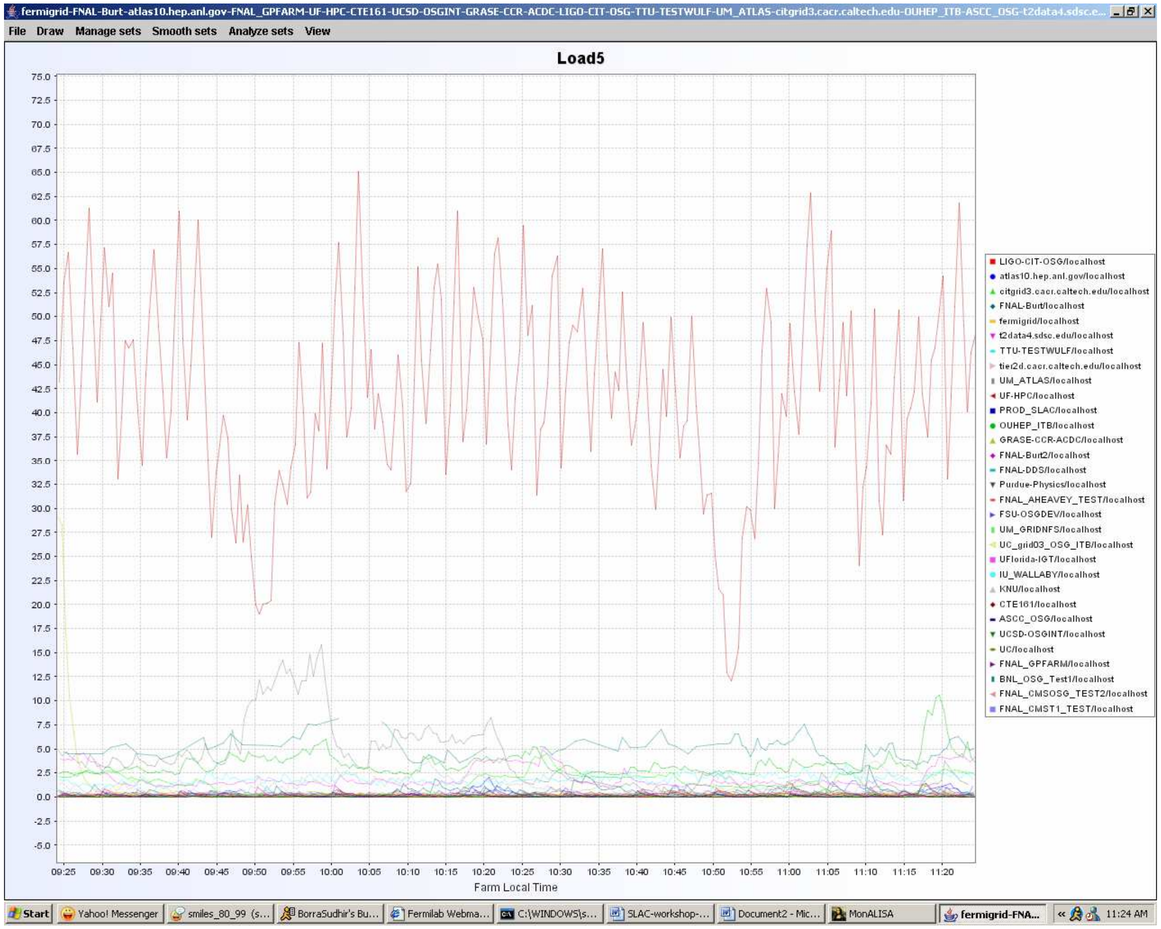

CMS FTP Transfer
Input (kbytes)
1000000
900000
800000
700000
600000
500000
400000
300000
200000
100000
0
Output (kbytes)
90000000
80000000
70000000
60000000
50000000
40000000
30000000
20000000
10000000
0
Feb 7, 00:00
Feb 14, 00:00
Feb 21, 00:00
Feb 28, 00:00
Mar 7, 00:00
Mar 14, 00:00
Mar 21, 00:00
Mar 28, 00:00
Apr 4, 00:00
Apr 11, 00:00
Apr 18, 00:00
Apr 25, 00:00
May 2, 00:00
May 9, 00:00
May 16, 00:00
May 23, 00:00
May 30, 00:00
GMT time
ASCC_OSG
atlas10.hep.anl.gov
BNL_OSG_Test1
citgrid3.cacr.caltech.edu
DUKE-HEP
fermigrid
FNAL-Burt
FNAL-Burt2
FNAL-DDS
FNAL_AHEAVEY_TEST
FNAL_CMST1_TEST
FNAL_GPFARM
fnpcg.fnal.gov
FSU-OSGDEV
GRASE-BINGHAMTON-ITB
GRASE-CCR-ACDC
IU_iuatlas
IU_WALLABY
KNU
LIGO-CIT-OSG
OUHEP_ITB
PROD_SLAC
Purdue-Physics
t2data4.sdsc.edu
tier2d.cacr.caltech.edu
TTU-TESTWULF
UBuffalo-CCR-ITB
UC_grid03_OSG_ITB
UERJ_HEPGRID
UF-HPC
UFlorida-IGT
UM_ATLAS
UM_GRIDNFS

trerCMS


ATLAS FTP Transfer
Input (kbytes)
550000
500000
450000
400000
350000
300000
250000
200000
150000
100000
50000
0
Output (kbytes)
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
Feb 7, 00:00
Feb 14, 00:00
Feb 21, 00:00
Feb 28, 00:00
Mar 7, 00:00
Mar 14, 00:00
Mar 21, 00:00
Mar 28, 00:00
Apr 4, 00:00
Apr 11, 00:00
Apr 18, 00:00
Apr 25, 00:00
May 2, 00:00
May 9, 00:00
May 16, 00:00
May 23, 00:00
May 30, 00:00
GMT time
ASCC_OSG
atlas10.hep.anl.gov
BNL_OSG_Test1
citgrid3.cacr.caltech.edu
DUKE-HEP
fermigrid
FNAL-Burt
FNAL-Burt2
FNAL-DDS
FNAL_AHEAVEY_TEST
FNAL_CMST1_TEST
FNAL_GPFARM
fnpcg.fnal.gov
FSU-OSGDEV
GRASE-BINGHAMTON-ITB
GRASE-CCR-ACDC
IU_iuatlas
IU_WALLABY
KNU
LIGO-CIT-OSG
OUHEP_ITB
PROD_SLAC
Purdue-Physics
t2data4.sdsc.edu
tier2d.cacr.caltech.edu
TTU-TESTWULF
UBuffalo-CCR-ITB
UC_grid03_OSG_ITB
UERJ_HEPGRID
UF-HPC
UFlorida-IGT
UM_ATLAS
UM_GRIDNFS


Number of CPUs in farms
550
500
450
400
350
300
250
200
150
100
50
Feb 7, 00:00
Feb 14, 00:00
Feb 21, 00:00
Feb 28, 00:00
Mar 7, 00:00
Mar 14, 00:00
Mar 21, 00:00
Mar 28, 00:00
Apr 4, 00:00
Apr 11, 00:00
Apr 18, 00:00
Apr 25, 00:00
May 2, 00:00
May 9, 00:00
May 16, 00:00
May 23, 00:00
May 30, 00:00
GMT time
ASCC_OSG
atlas10.hep.anl.gov
BNL_OSG_Test1
citgrid3.cacr.caltech.edu
DUKE-HEP
FNAL-Burt
FNAL_CMSOSG_TEST2
FNAL_CMST1_TEST
FNAL_GPFARM
fnpcg.fnal.gov
FSU-OSGDEV
IU_iuatlas
IU_WALLABY
KNU
LIGO-CIT-OSG
OUHEP_ITB
Purdue-Physics
tier2d.cacr.caltech.edu
TTU-TESTWULF
UC_grid03_OSG_ITB
UERJ_HEPGRID
UERJ_HEPGRID_ITB
UF-HPC
UFlorida-IGT
UIOWA-ITB
UM_ATLAS
UM_GRIDNFS


Farm usage
Number of busy nodes
0
25
50
75
100
125
150
175
Number of free nodes
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
Feb 7, 00:00
Feb 14, 00:00
Feb 21, 00:00
Feb 28, 00:00
Mar 7, 00:00
Mar 14, 00:00
Mar 21, 00:00
Mar 28, 00:00
Apr 4, 00:00
Apr 11, 00:00
Apr 18, 00:00
Apr 25, 00:00
May 2, 00:00
May 9, 00:00
May 16, 00:00
May 23, 00:00
May 30, 00:00
GMT time
ASCC_OSG
atlas10.hep.anl.gov
BNL_OSG_Test1
citgrid3.cacr.caltech.edu
DUKE-HEP
FNAL-Burt
FNAL-Burt2
FNAL-DDS
FNAL_CMSOSG_TEST2
FNAL_CMST1_TEST
FNAL_GPFARM
fnpcg.fnal.gov
FSU-OSGDEV
IU_iuatlas
IU_WALLABY
KNU
LIGO-CIT-OSG
OUHEP_ITB
Purdue-Physics
tier2d.cacr.caltech.edu
TTU-TESTWULF
UC_grid03_OSG_ITB
UERJ_HEPGRID
UERJ_HEPGRID_ITB
UF-HPC
UFlorida-IGT
UIOWA-ITB
UM_ATLAS
UM_GRIDNFS



Load on Master Nodes sampled every 5 minutes.