Field of Science Reporting in the OSG

Overview

Most of the reporting in the OSG groups its data in one of two ways – by organization (VO) that submitted the jobs or by the site that ran the jobs. This is a natural approach, because it groups jobs either by producer (the VO) or the consumer (the site).

We are working on a third method of groupings: field of science. Grouping data by science has immediate strengths:

1. External parties don't recognize VOs by their names, especially smaller VOs. Classifying by science allows us to cast the data in more familiar terminology.

2. Knowing the science provides the OSG with feedback on how far they are spreading outside of the field of HEP, which has been the primary driver this far.

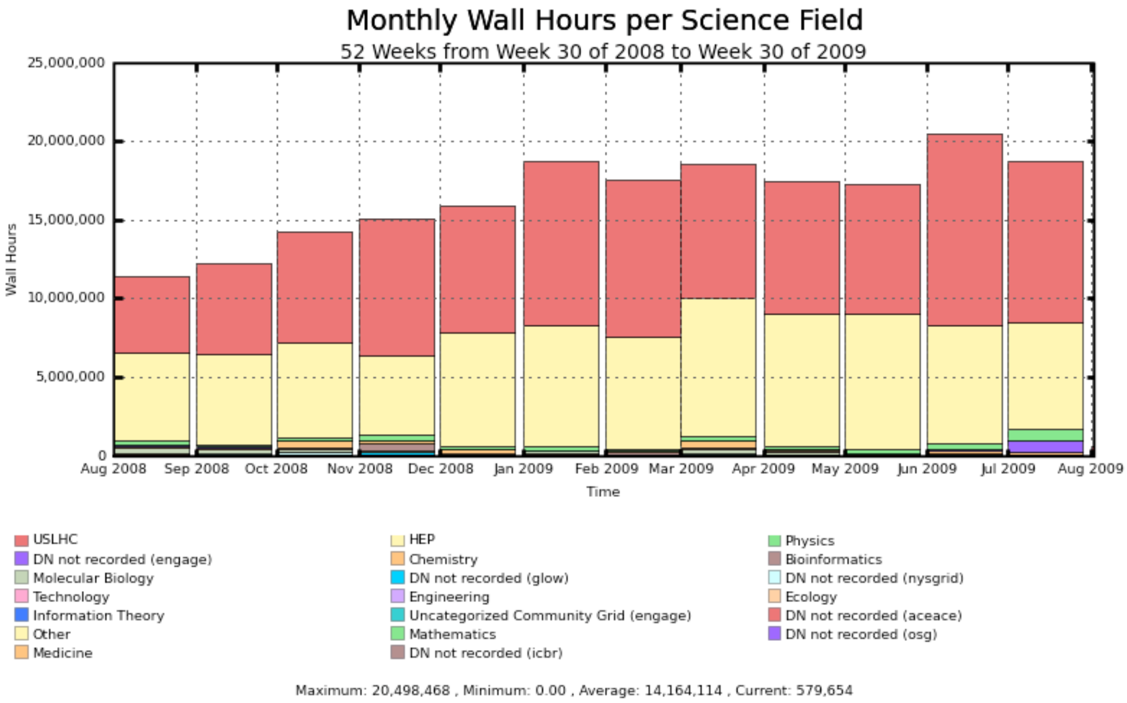
This document covers the current progress and results in reporting accounting data by field of science.

While the OSG has a thorough accounting system with Gratia, several critical pieces of data had to be developed to complete this project. The largest hurdle was handling community grid VOs. These organizations have many users with multiple science domains; unlike experiment VOs, it was not as simple as labeling the entire VO as being in one domain.

Results

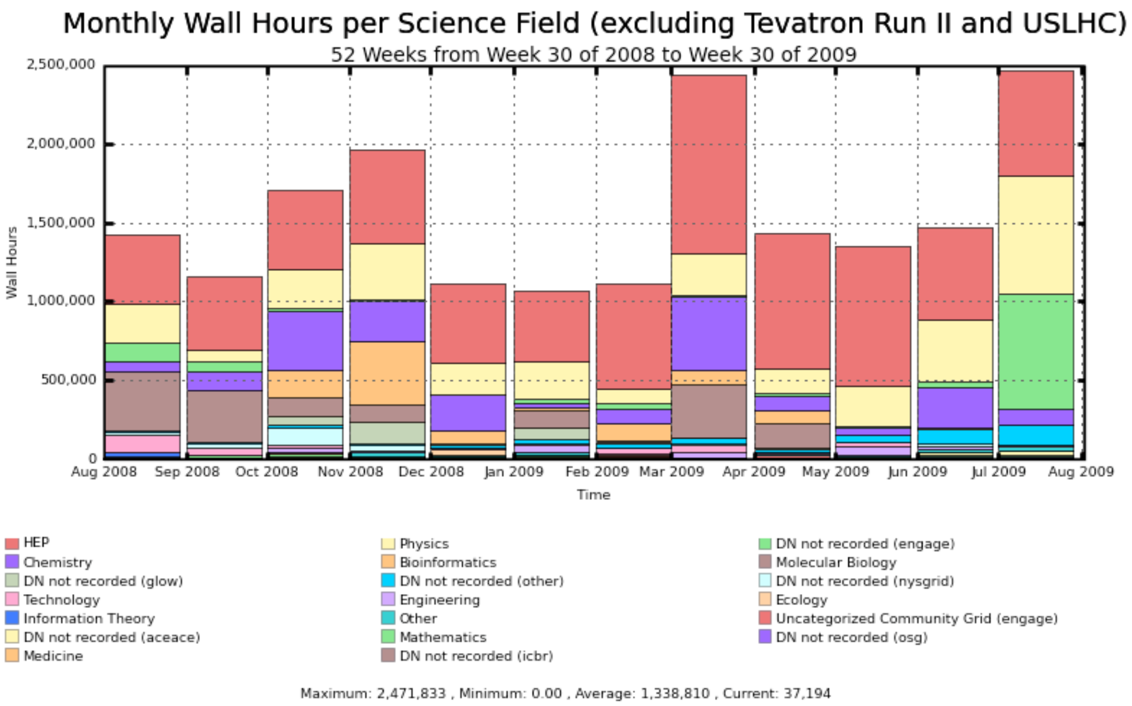
The graph below shows the number of wall hours spent on jobs by field of science[1].

[1] http://t2.unl.edu/gratia/bar_graphs/monthly_field?starttime=2008-08-01%2000:00:00&endtime=2009-08-01%2023:59:59



Because USLHC (CMS and ATLAS) and Tevatron Run II (D0 and CDF) experiments are so dominant, it helps to remove those four VOs[2]:

[2] http://t2.unl.edu/gratia/bar_graphs/monthly_field?starttime=2008-08-01%2000:00:00&endtime=2009-08-01%2000:00:00&exclude-vo=cms|atlas|cdf|dzero|unknown



Several entries in the legend are worth explaining, as they point out the limitations of our data collection method:

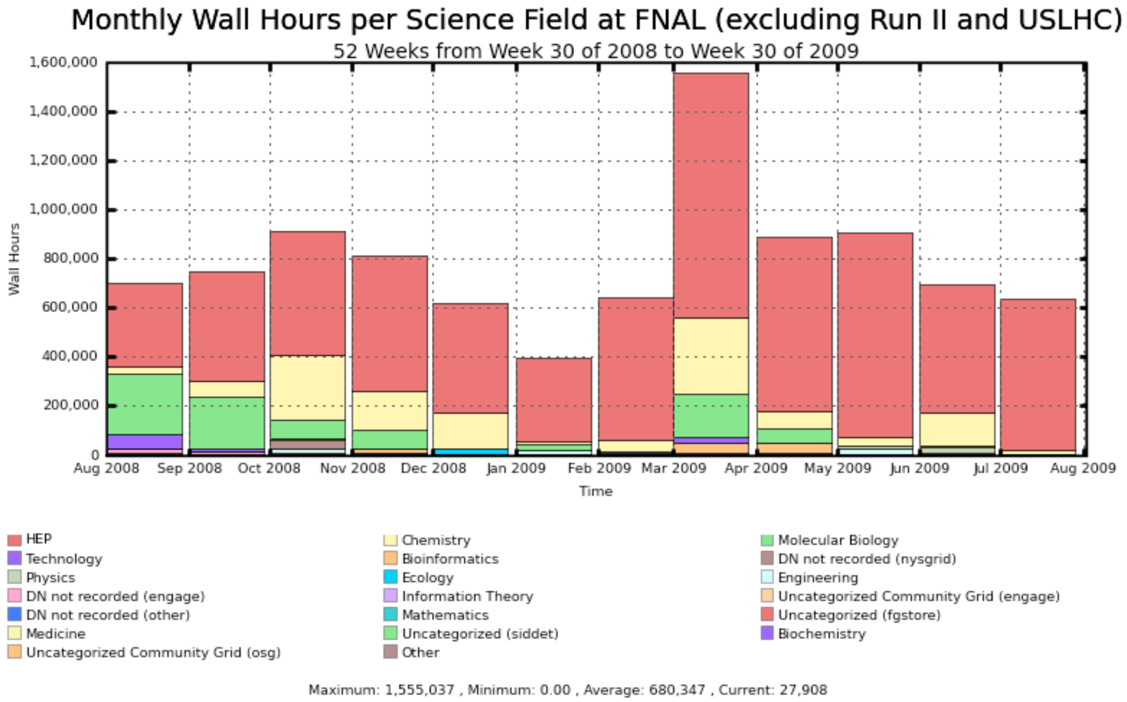
- **Other**: Any entry whose usage is too small to display on the graph gets aggregated into this catch-all entry.

- **DN not recorded (VO name)**: Not all sites report the certificate's Distinguished Name (DN) associated with a job, due to misconfiguration, complex site issues, or historical data that was reported prior to Gratia having this capability. This entry corresponds to jobs without reported DNs whose VO does not have a single science classification. These entries are due to missing information in the job record, and it is not possible to retroactively correct past data; any site has the ability to report this correctly.

- **Uncategorized Community Grid (VO name)**: This is usage in a community grid which has a DN associated it, but the VO has not yet determined what that user's science community is. When the VO has made that determination, we will be able to label this historical data.

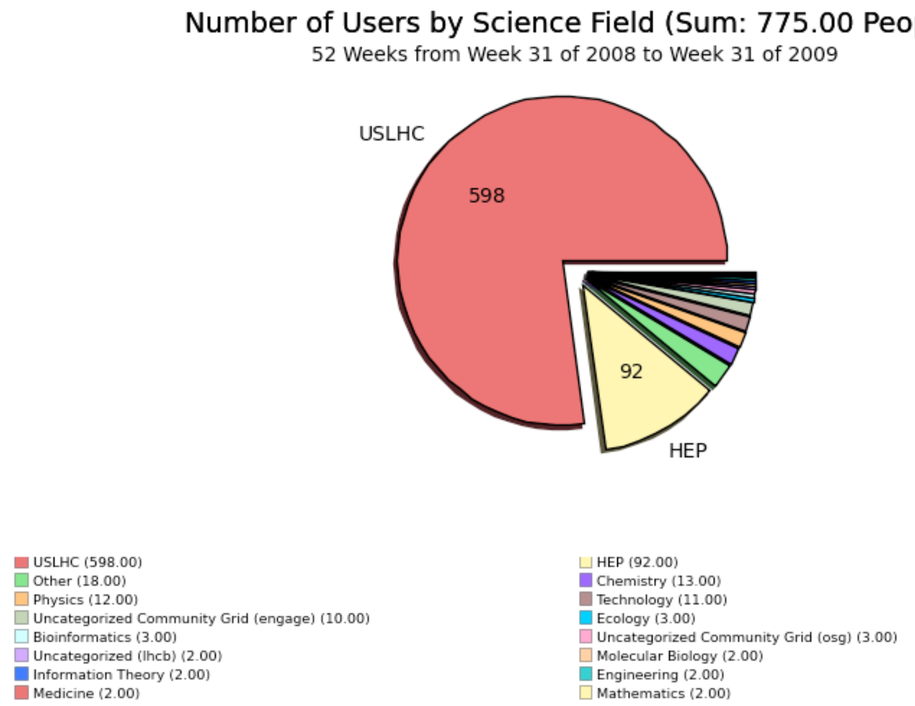
- **Uncategorized (VO name)**: This is usage in a VO that has not been classified at all. Usually, this signals a VO that runs at a site but has not registered with the OSG.

One requested use of this data is to classify the science being performed at a single site (i.e., so the executives know what areas they are contributing to). Below, we have an example of the science being done at the FNAL facility[3].



Below is the number of “significant” users recorded per science field[4]. Here, we define a user to be “significant” if they’ve used more than 1000 wall hours.

[3] http://t2.unl.edu/gratia/bar_graphs/monthly_field?starttime=2008-08-01%2000:00:00&endtime=2009-08-01%2000:00:00&probe=fnal.gov&exclude-vo=cdf|dzero|atlas|cms|unknown
[4] http://t2.unl.edu/gratia/pie_graphs/science_field_user_count



Method of Data Collection

All OSG VOs are classified according to their science field or "community grid" if their users can be classified under multiple fields. For these community grids, we have asked that they self-classify their users according to science field. We have developed a simple CSV format (so it can be edited in Microsoft Excel) that allows the community grids to classify a certificates Distinguished Name (DN), corresponding to a single user, or a FQAN, corresponding to a group of users. They then host this CSV file on their own servers and update it as users come or go[5]. Ideally, this would allow Metrics and Measurements to not have to maintain any data centrally. For this report, we have used a separate CSV to classify users for VOs that are not yet up to speed.

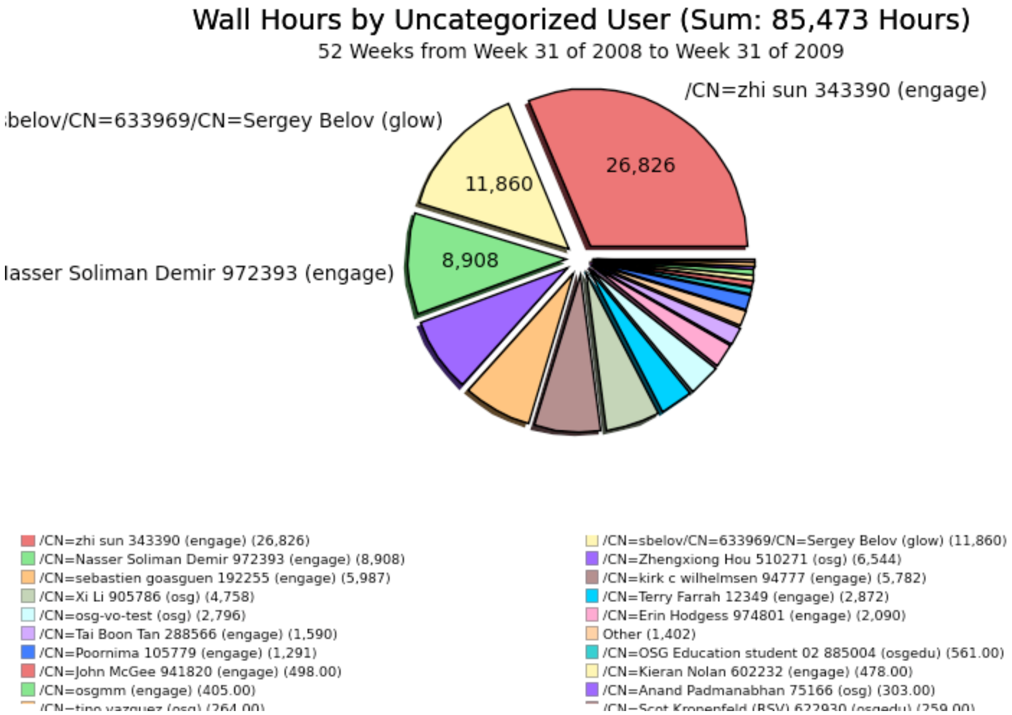
Each job record has an associated VO and may also have a DN and FQAN. The information we have received from VOs and OIM is applied to each record; it is applied first to the DN, then FQAN, then VO. If information is missing, we may apply one of three labels: DN not recorded, Uncategorized, or Uncategorized Community Grid. These labels are explained above.

The two largest (in terms of wall hours) community grids Engage and NYSGrid are currently participating; we expect the OSG VO to participate in the near future.

We provide feedback to the VOs to help them know which users have not yet been classified, sorted by wall hours. The pie graphs below show this information for the

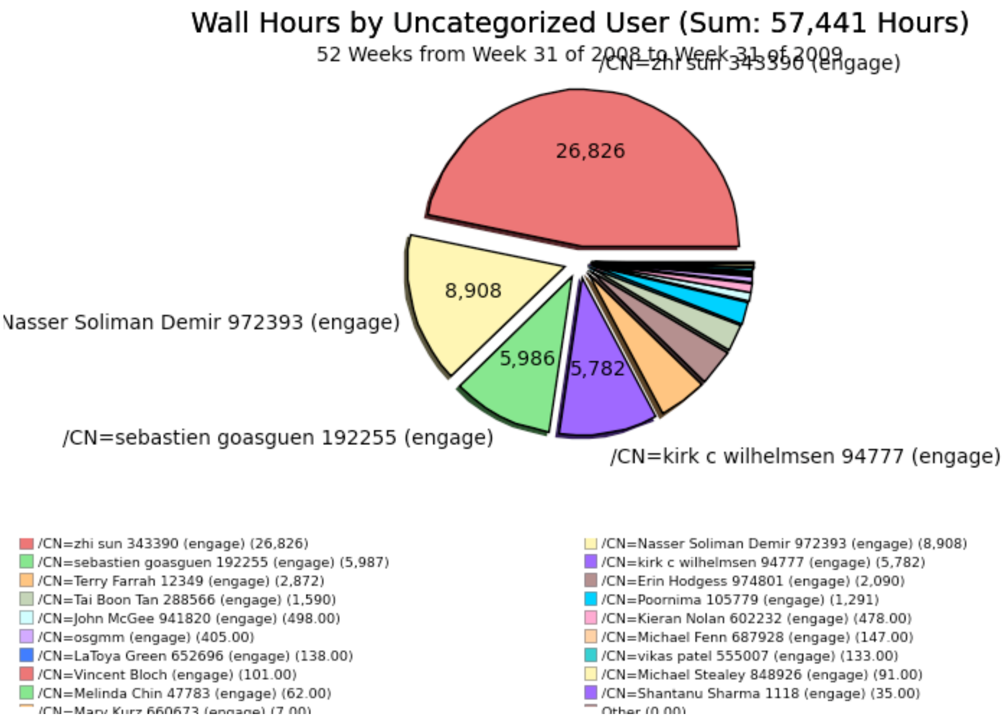
[5] For an example of a VO's CSV, see the engage one: http://engage-central.renci.org/engage-sciences.csv

whole OSG[6], then just for Engage[7]. Currently, all the uncategorized usage adds up to about 10 CPU years; once we remove the top 5 users, uncategorized usage will be about 3 CPU years.



[6] http://t2.unl.edu/gratia/pie_graphs/unclassified_users

[7] http://t2.unl.edu/gratia/pie_graphs/unclassified_users?vo=engage



VOs were given the following initial list of science fields to use to classify their users (but were able to add fields as needed):

- Astronomy
- Biochemistry
- Bioinformatics
- Biophysics
- Botany
- Cellular Biology
- Chemistry
- Community Grid
- Computer Science
- Earth Sciences
- Ecology
- Engineering
- HEP
- Information Theory

- Logic
- Mathematics
- Medicine
- Microbiology
- Molecular Biology
- Physics
- Physiology
- Statistics
- Technology
- USLHC
- Zoology

Some VOs did not fill in this information at registration time (many VOs registered before it was possible to fill in field of science). For these VOs, and for inactive VOs, we have put in a field of science to the best of our ability. VOs that are still active will be contacted and asked to review the data we filled in for them. The active VOs whose data we altered are:

- CIGI
- CompBioGrid
- Engagement
- Fermilab/Astro
- Fermilab
- Fermilab/Minerva
- Fermilab/Mu2e
- Fermilab/Nova
- Fermilab/Numi
- Fermilab/Patriot
- Fermilab/Test
- GLOW
- GPN

- GROW
- i2u2
- JDEM
- mariachi
- MIS
- nanoHUB
- Ops
- OSG
- OSGEDU