Analysis of job submissions through the EGEE Grid

Mona Aggarwal, David Colling, Janusz Martyniak,
A. Stephen McGough, Gidon Moont and Olivier van der Aa
Imperial College London

Overview

The Grid as an environment for large scale job execution is now moving beyond the prototyping phase to real deployments over national and international scales providing real computational cycles to application scientists who are using it for real scientific applications. These real deployments are highlighting characteristics of these Grids which could not have been predicted before major deployment. In order to better understand these characteristics a full analysis of these Grids needs to be performed. In this work we analyse trace logs of over 70 million jobs collected from jobs executed through the Enabling Grids for EscienceE (EGEE) Grid. A large worldwide Grid consisting of over 41,000 CPU's and 5PB of online disk storage spread over 240 institutions from over 45 countries. Users are members of different Virtual Organizations (VO's) based on shared projects and/or geographical location.

Performance Metrics

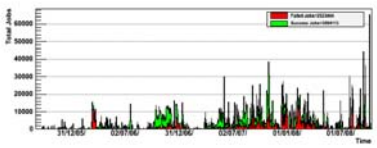
We use the following metrics in this work:

- **Job Submissions** in a given day.
- **Cumulative job submissions** in a given day.
- **Active Users per VO** – number of users from a given VO that have been active on a given day.
- **Job Hours** – total number of hours consumed by a VO on a given day.
- E_{RAW} = Total number of successful jobs in a day / Total number of jobs in that day
- EUSER = Time job is running . / Time Job is in system

BIOMED

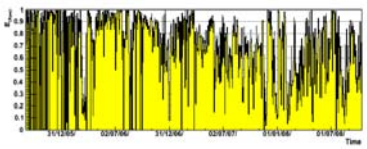
Job submissions vs time

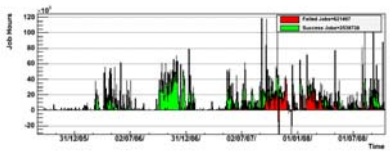
Three main periods of activity – Grand Challenges. Many failed jobs at other time (testing).



E_{RAW} vs time

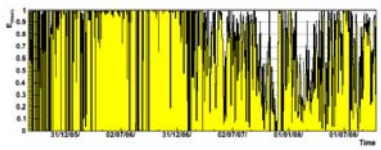
During Challenges RAW efficiency is good, when job load is low RAW efficiency is variable – matches testing.





Job Hours vs time

Hours of successful work are high in comparison to failed hours. Failures happen fast (as desired). General increase in hours used.

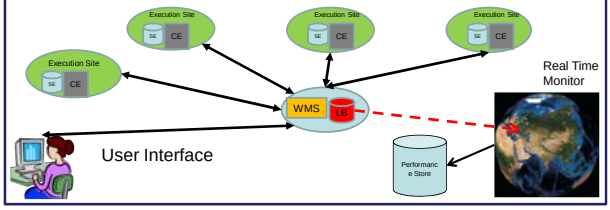


E_{USER} vs time

User efficiency is good during challenges and inconsistent in-between – due to low job numbers.

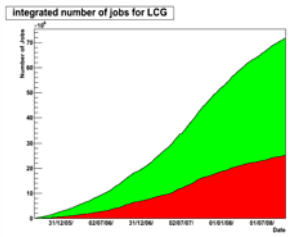
Data Collection

Imperial College London (High Energy Physics Group) have been developing tools for monitoring the state of jobs on the Grid. The culmination of this work is the Real Time Monitor (RTM http://gridportal.hep.ph.ic.ac.uk/rtm/). The RTM queries the LB's within the Grid directly and makes this information available to the end user in a graphical format. We have also been logging this information since September 2005, collecting statistics on over 70 million jobs.



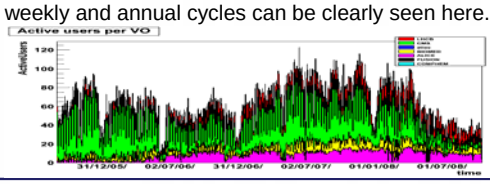
Growth of the EGEE Grid

The Cumulative job submissions graph shows a greater than linear increase since September 2005.



VO User Activity

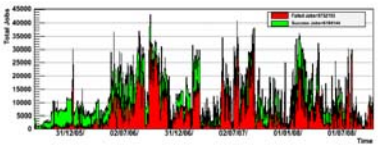
VO's have between 43 and 1600 members though no more than 120 users were active on a given day, In general users will naturally interleave Grid work with other work. The weekly and annual cycles can be clearly seen here.



LHCb

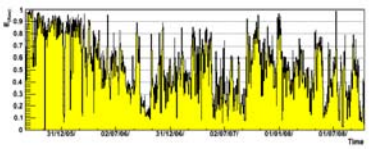
Job submissions vs time

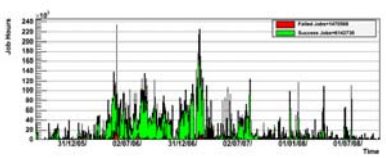
Production runs mid 2006 and 2007. They use pilot jobs which fail quickly if resource is wrong or no waiting job.



E_{RAW} vs time

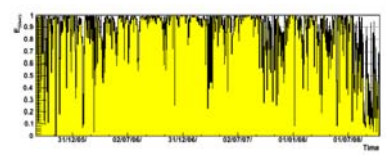
Pilot jobs make the RAW efficiency low. Especially when in production runs.





Job Hours vs time

Though many failed pilot jobs these take very little time so very little time is spent running failed jobs.



E_{USER} vs time

User efficiency is very good as jobs which run do so quickly and in general there is a high job load from this VO.