Contribution ID: **84** Type: **Sessions contributions**

Towards harmonized workload management for the biomed VO with DIRAC

Thursday, 22 May 2014 14:00 (25 minutes)

biomed has been the most active life-science EGI VO for the last 3 years, representing 77% of the normalised CPU time consumed for life sciences on the infrastructure. For various reasons, workload management was never centrally coordinated in the VO, which resulted in two issues: (i) central monitoring is lacking; (ii) computing resources are not efficiently exploited. To address these issues, we are investigating the adoption of DIRAC as a central workload management solution in the VO. In addition to central monitoring, DIRAC provides an efficient pilot-job mechanism to balance the load among computing elements. However, much remains to be done for biomed to be able to exclusively rely on DIRAC for workload management. In particular:
* One should make sure that DIRAC can actually be used by all existing users, tools, frameworks and portals in the VO.
* User-level documentation should be widely available and maintained.
* A production-level DIRAC instance able to support the whole VO activity should be deployed and operated.
Once this is addressed, we are confident that using DIRAC at a VO-level would globally improve workload management, as already reported by large international collaborations. That would also open other perspectives in terms of data management and access to different resource types.

Wider impact and conclusions

Harmonizing workload management enables better monitoring and policy enforcement in a VO, as demonstrated by High-Energy Physics. DIRAC has been successfully used by large international collaborations (e.g. LHCb) and some groups within biomed. Beyond workload management, consistently using in the biomed VO would open a few perspectives:
* DIRAC's data management can handle cleanup and "dark data", two issues very problematic in biomed that have to be treated at the VO level.
* DIRAC can handle several types of computing resources, for instance cloud and single clusters.

URL(s) for further info

http://lsgc.org

Description of work

Thanks to the large support provided by National Grid Initiatives, the biomed international VO has been the most active life-science EGI VO for the last 3 years, representing 77% of the normalised CPU time consumed for life sciences. Operations in biomed are supported by teams of shifters monitoring the core services through Nagios. Users are free to use or any software solution to access these resources. As a consequence, a few issues hinder workload management :
* Central monitoring is lacking. No central service is available for job monitoring. Consequently, user mistakes and operational issues such as unexpected job failures can only be detected once reported by users or site administrators.
* Efficiency is only fair. Computing resources are not optimally used; it happens that a few sites are flooded by jobs while others are mostly idle.

Some solutions have been explored to handle these issues. For instance, the VAPOR portal provides white lists of well-performing sites, and the DIRAC pilot-job system is available. DIRAC has been provided by France-Grilles since 2012. In 2013, it handled 38% of the CPU time consumed by the VO.Adopting DIRAC as the single workload management solution in the VO would have several advantages for workload management. In particular:
* DIRAC provides central monitoring: jobs and their statuses can be monitored and operational issues are

quickly detected.
* DIRAC provides a pilot-job system balancing the load among computing elements very efficiently.

However, much remains to be done for biomed to be able to exclusively rely on DIRAC for workload management. In particular:
* One should make sure that DIRAC can actually be used by all existing users, tools, frameworks and portals in the VO.
* User-level documentation should be widely available and maintained.
* A production-level DIRAC instance able to support the whole VO activity should be deployed and operated.

Primary authors: TSAREGORODTSEV, Andrei (CNRS); MICHEL, Franck (CNRS); PANSANEL, Jerome (CNRS); Mr MONTAGNAT, Johan (CNRS); Dr GRACIANI DIAZ, Ricardo (University of Barcelona); GLATARD, Tristan (CNRS)

Presenter: GLATARD, Tristan (CNRS)

Session Classification: DIRAC Virtual Research Environment pilot for EGI

Track Classification: Virtual Research Environments, gateways and workflow engines (Track Leaders: J. Montagnat, G. Sipos)