LBNE Feedback for the "Requirements for the Batch Submission System for Frontier Experiments"

FIFE Workshop

June 4th, 2013

Maxim Potekhin
Brookhaven National Laboratory

potekhin@bnl.gov

Overview

- In terms of computational resources scale, LBNE will eventually be on the high end of the spectrum. What is mentioned as "nice to have" in the feedback from Minerva and MicroBooNE is likely to be "must have" in LBNE (e.g. multiple sites etc). We must make sure our future needs are aligned with the rest of IF and their requirements.

- The document may benefit from
  - Broadening the scope of the engineering issues involved (such as being more specific with regards to what "off-site" submission can be)
  - Providing a more detailed description of specific stages in implementation (clearly some of the requirements are near-term; what about more detail regarding the timeline?)

- The topic of Monitoring (specifically at the job level) may deserve more focus in this paper

- Does not necessarily belong in the requirements, but nice to know and/or have a summary, possible in a separate write-up:
  - An overview of the computing requirements and the ways they are being met, in disciplines such as HEP and Nuclear Physics, so as to leverage existing experience

The Big Picture

It appears from the document the long-term requirements (to the extent presented) are not really unique to the Intensity Frontier, and are mostly aligned with what other HEP experiments have been facing in recent past. Important aspects from this category include submission of jobs to the Grid, management of workload across multiple sites, resource allocation and prioritization etc.

From experience of the past decade, successfully meeting these challenges involves a more systemic look at the requirements, and an integrated engineering solution. In many cases, the solution presents itself as what is termed "Workload Management System" (WMS). Examples include DIRAC, glideinWMS and PanDA. Work is underway to advance new HTCondor-based systems such as Bosco, having its own interesting set of capabilities.

In most cases, a high degree of transparency to the end users, optimal resource allocation and considerable scalability have been achieved. "Jobsub" is an example of a successful overlay with glideinWMS – but still essentially Condor.

Federation of resources achieved by WMS, provided it is accompanied by effective and comprehensive monitoring, opens the way for efficient operational support, where the effort can be shared among participating institutions and thus provide better quality and more responsive coverage.

WMS

Typically, the WMS takes care of the following (in varying degrees, depending on implementation):

- Deployment of users' jobs on participating grid sites (cf. "seamless way" in the requirements), while providing a uniform interface to the user, regarding of whether the local batch system is used or the job is to be executed on a remote CE.
- Authorization and authentication, often based on X.509
- Brokerage, i.e. matching of jobs submitted by users to the actual computing resources based on various criteria, such as data location, hardware characteristics, resource allocation etc
- Utilization of VM and Cloud-based resources
- Prioritization, including dynamic adjustment of jobs' priorities
- Transparent monitoring – again, without regard of location of the job. In some cases, access is provided to individual and detailed log files generated by the user's jobs, as well as web-based summaries. Job status, error conditions and location is made available to users and operators. This is one of the more powerful features of WMS which everybody thinks is useful. Integrated view of sites' performance is also provided
- Data handling. This item varies across different WMS, and can be as simple as more or less native Condor stage-out of the data (in glideinWMS) or as sophisticated as DDM in ATLAS PanDA.

Conclusions

We in LBNE Software and Computing Organization would like to see if there is interest among the stakeholders to consider the goal of “batch submission” from a more general viewpoint of a Workload Management System which utilizes Grid computational resources.