Interim report on status of VO ability to use OSG: DZERO, Fermilab, MiniBooNE, GADU and SDSS/DES

October 26, 2006

0. Introduction

This is a brief status report on work being done to enable several VOs to use as much as possible of the wider OSG resources for upcoming needs. Section 1 describes the specific needs of each VO where applicable; section 2 shows a summary table and explains various aspects of how the data were collected; section 3 describes various problems encountered and steps being taken or to be taken to address them.

1. Specific VO characteristics and needs.

DZERO

- 1500-2000 CPUs for a period of about 4 months.
- Outgoing network access from the worker nodes.
- Worker nodes should have about 6GB of scratch storage space per job.
- 4 TB of total disk space to store input and output files distributed among the participating sites.
- Desirable to have at least 1 TB of disk cache connected with 1 Gbit link to all CPUs.

Fermilab

No specific needs.

GADU

- Of order 1 month of around 1000 job slots.
- Software on OSG_APP, pushed.
- Automated job dispatch system to pre-verified sites upon which software has already been installed.
- Jobs 2-6 hours long each
- Heavy access to file-DB (about 3GB total) throughout jobs. Many jobs running simultaneously can cause strain on some fileservers and (sometimes) cause trouble for the headnode. Working on fixing (see notes on OSG_WN_TMP); meantime, identify and throttle jobs on susceptible sites.

SDSS / DES

- Require outgoing network connections from worker nodes.

MiniBooNE

- Some jobs (data processing, not MC) require outgoing network connections from worker nodes.
- Ongoing heavy MC production: jobs range from 10 mins (refit) to 12 hours (full MC run).
- Software and global auxiliary data installed in OSG_APP by first job to need it (locked pull operation): about 1.2GB.
- Large amount (up to 4GB) of per-job input data required for full MC jobs.

2. Progress Summary

See site summary table for fine details.

Legend and explanation

A red block in place of data in the "site information" section indicates that those data were not obtainable. Some sites were not listed in Mona Lisa; gridcat and LDAP were also utilised as sources for job slot information but the information was either unreliable or misisng.

A mid-grey block in place of data in the VO comments section indicates that the site did not advertise itself as supporting that VO.

BASIC_PASS: site passed basic tests (ping, globus-job-run (printenv and df), gridftp push and pull to head node);

FTP_FAIL: gridftp push/pull failed;

JOBMAN_FAIL: printenv job failed apparently for some other reason than no authorization;

NO_AUTH: authorization failure;

READY (GADU only): tests passed, software is installed and awaiting production start;

RUNNING: this site is being utilized for production jobs;

X (SDD/DES): site is advertised as available: VO-specific test results not yet known;

GOC Ticket #2644: see details of ticket on GOC site or brief details below.

3. Details.

Some interesting things were observed during this exercise: they are presented below in no particular order:

- The site lists were obtained from VORS using wget and scripting. It seems that the list of sites advertising themselves as available to a particular VO may not be completely reliable (see GADU / UIOWA-OSG-PROD). Subsequent test runs will obtain the master OSG list rather than relying on the VO-specific lists).
- Many sites claim to support a particular VO but do not; conversely, some sites (eg FNAL) **should** support a particular VO but do not (cf DES, DZERO).
- Some sites appear to be mis-configured, in that the OSG variables do not appear to be available to a job (at least to a globus-job-run).
- A reasonable way to solve GADU's NFS access problems is not actually possible under the

current OSG setup. OSG_WN_TMP is supposed to be a job-specific area on the worker node which should (actually, may be automatically) cleaned up at the end of a job. Ideally, there would be an area (OSG_WN_DATA, for example) on the worker node which followed the rules of OSG_DATA (not guaranteed to be available to subsequent jobs but not explicitly cleaned up) where the 3GB file-DB could be deposited and accessed locally.

- BNL_ATLAS_{1,2} in particular appears to use a mapping scheme which causes particular problems for me as a user in multiple VOs. Briefly, it appears that the first time the system sees a particular DN, it looks in VOMS and finds the first VO of which this user is a member, permanently assigns a pool username (eg grid5002) and sets the GID based on the ascertained VO. It does not use voms-proxy info, and users have different UIDs (unlike eg FNAL which maps the VO from the incoming proxy to a single group user, like miniboone). There are many consequences of this type of site configuration which are problematic:
    - The pull-style of software installation in OSG_APP becomes tricky and/or insecure (need to have world-writable areas or know to do newgrp at the beginning of a job).
    - Both testers like myself and bona fide physics users who are members of multiple VOs will find it difficult if not impossible to operate successfully in all their capacities.
- Fermilab has recently moved to requiring voms-proxies, not just grid-proxies. This may be a problem for DZERO; this is currently being investigated.
- Over the next days and weeks, contact will be made with problematic sites (by direct negotiation in preference to GOC tickets) to resolve these issues and put more green on the summary table.