

Tier3 Support WG: GangaRobot and HammerCloud Status

Dan van der Ster, CERN IT-ES-DAS

*ATLAS T2/3 Workshop @ OSG All-Hands Meeting
9 March 2010*



Overview of GangaRobot and HammerCloud

- *GangaRobot* (GR) and *HammerCloud* (HC) are automated tools used by ATLAS to:
 - perform frequent **functional tests** of distributed analysis jobs (used for example to validate the sites)
 - run infrequent distributed analysis **stress tests** (used for example to commission a site or evaluate configuration changes)
- GangaRobot: <http://gangarobot.cern.ch>
- HammerCloud: <http://gangarobot.cern.ch/hc/>



DDM Awareness

- DDM-aware and non-DDM-aware Tier3s will need to be treated differently.
- DDM-aware T3's will work out-of-the-box for GR and HC, so no discussion is needed.
- Non-DDM-aware T3's (i.e. Panda w/ Batch Pilot Submitter) will be tricky to get working with GR and HC



GR and HC for "Panda" Tier3s

- When GR/HC submit a job, they need to know:
 - which athena release: we can suggest a standard testing release
 - which input data:
 - datasets available at a site
 - files within the datasets available at a site
 - PFN of the file at the site
- pathena/Ganga (around which GR/HC are built) require as an argument at submission time a pfnList
 - a file containing the list of PFN's to process at the site.
- GR/HC would need to know such a list for each site to be able to run tests



How to get the pfnList for GR/HC tests

- We have at least two options for getting the pfnList for GR/HC:
 - A) Each site submits a text file which is stored by GR/HC:
 - allows us to make use of the data which is already at a site
 - only sites having submitted a pfnList will be testable
 - consistency will be difficult/impossible to maintain
 - this sounds labour intensive
 - B) If files can be at the same physical path at all or many sites, then we can agree on a standard set of datasets to be tested:
 - i.e. all atlas data at /xroot/atlas/<datasetname>/.../<filename> (where '...' is an agreed upon function of the datasetname)
 - We agree on a set of GR datasets DS1..DSn which must be at all sites to be tested.
 - This way GR/HC can trivially generate the pfnList for all sites.
 - For GR (functional test) we only need a few files to test.
 - For HC (stress test) we usually require many GB or a few TB of data for a test. Impractical? Impossible?



Summary

- Both (A) and (B) make some sacrifices.
- Is there another approach to solve this problem? e.g. some kind of trivial central cataloging of the data at the Tier3s?

