
Subject: Re: Data Challenge

Posted by [Dan Protopopescu](#) on Thu, 03 Apr 2008 10:34:42 GMT

[View Forum Message](#) <> [Reply to Message](#)

Dear all,

I would like to thank everybody for participating in this discussion. There are several that showed up and I will address herein:

- 1) Dates of the DCs
- 2) Nightly builds on Grid
- 3) Proposed test jobs/macros
- 4) Storage of the outputs

Here below is my opinion on these issues. Please use the forum to reply so that we have an organized thread. If you have no access, register or ask someone else to post for you.

1) The Grid Data Challenges will happen at dates to be decided independently and to ensure this objectivity I propose that the dates be set by our Production Manager (PM), Paul Buehler, after some consultation with both the Grid Coordinator (Dan) and Software Development Coordinator (Johan). The next date is April 17 (no changes accepted) but after that I hope the proposed PM scheme will be applied.

2) One of the major points to be understood about Grid is that it is not a testbench. It is a "massive computations infrastructure", designed for large-scale, standardized jobs. Although nightly builds on several platforms is an extremely useful tool for developers to spot early compilation problems, it would be a misuse of the grid. The software installed on the grid is supposed to be a stable version, which has been tested already on the platforms existent on the grid sites. Testing and feedback would happen at installation time. Florian explained this issue better than me.

3) The basic plan for the next data challenge is to count jobs and produce statistics like job success rate, job site distribution, time per 1000 jobs. I propose the tests to be done both with a generic job and a PandaRoot job in order to decouple the various requirements. Having a macro that simulates some tracks in the EMC or producing some real physics is almost irrelevant for this data challenge. However, I enthusiastically embrace the idea of running some physics from which someone can collect, verify the results and gain an extra benefit. If you have such macro, let's use it!

My initial plan is this:

- 10x100 subjobs generic (site availability)
- 1x1000 subjobs generic (broker optimization)
- 1x1000 PandaROOT macro #1
- 1x1000 PandaROOT macro #2
- 1x1000 PandaROOT macro #1 or #2 with all output to Glasgow SE
- 1x1000 PandaROOT macro #1 or #2 with output to local SEs

Please feel free to add to this list and let's discuss the benefits of these tests.

Package testing can not be part of the data challenge because of time constraints. A check of the software installation is part of the preparations taken individually by the site admins.

4) The outputs of the DC test jobs will go physically to local and central SEs and that is part of the challenge. In the file catalogue, the output can be collected as you see fit, in case you would like to use the results later.

Cheers,
Dan
