Contribution ID: **116** Type: **oral presentation**

Data Management in LHCb: consistency, integrity and coherence of data

Thursday 10 May 2007 11:40 (20 minutes)

Describe the scientific/technical community and the scientific/technical activity using (planning to use) the EGEE infrastructure. A high-level description is needed (neither a detailed specialist report nor a list of references).

The Large Hadron Collider (LHC) at CERN will start operating in 2007.
The LHCb experiment is preparing for the real data handling and analysis via a series of data challenges and production exercises. The aim of these activities is to demonstrate the readiness of the computing infrastructure based on WLCG (Worldwide LHC Computing Grid) technologies, to validate the computing model and to provide useful samples of data for detector and physics studies.

Report on the experience (or the proposed activity). It would be very important to mention key services which are essential for the success of your activity on the EGEE infrastructure.

The LHCb policy on these issues has been addressed towards an investment in resources targeting
the minimization of the number of occurrences involving data corruptions, data missing, data incoherence and inconsistencies among Catalogues and physical storages, both through safety measures at data management level (failover mechanisms, check sums, roll back mechanisms) and through expensive background checks. The data integrity and the consistency checks activity are presented here. The goal of this activity is to be able to maintain a consistent picture of the main catalogues (Bookkeeping and LFC) and the Storage Elements, primarily among them, and at a second order with the computing model.
While reducing actively the number of these interventions still represents the main goal of the DMS in LHCb, the outcome of these checks represents also a lucid evaluation of the quality of service offered by the underlying Grid infrastructure.

With a forward look to future evolution, discuss the issues you have encountered (or that you expect) in using the EGEE infrastructure. Wherever possible, point out the experience limitations (both in terms of existing services or missing functionality)

The planned activity on data integrity, consistency and coherence in the Grid is addressed for the
development, in a near future, of a generic tool suite able to categorize, analyze and systematically cure the disparate problems affecting data management. The advantages are: the efforts made to solve immediate problems can be embedded in more generic and higher level tools; and fixes to some problems can be applied to DIRAC as well to avoid repetitions of problems.

Describe the added value of the Grid for the scientific/technical activity you (plan to) do on the Grid. This should include the scale of the activity and of the potential user community and the relevance for other scientific or business applications

DIRAC (Distributed Infrastructure with Remote Agent Control) is the gateway to WLCG. The Dirac Data Management System (DMS) relies on both WLCG Data Management services (LCG File Catalogues, Storage Resource Managers and File Transfer Service) and LHCb specific components (Bookkeeping Metadata File Catalogue). Although the Dirac DMS has been extensively used over the past years and has proved to achieve a high grade of maturity and reliability, the complexity of both the DMS and its interactions with numerous WLCG components as well as the instability of facilities concerned, turned frequently into unexpected problems in data moving and/or data registration. Such problems make it impossible at all times to have a coherent picture of experimental data-grid across various services involved.

Author: BARGIOTTI, Marianne (European Organization for Nuclear Research (CERN))

Presenter: BARGIOTTI, Marianne (European Organization for Nuclear Research (CERN))

Session Classification: Data Management

Track Classification: Data Management