

ARDA Dashboard Data Management Monitoring

Thursday 10 May 2007 12:00 (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 Atlas DDM (Distributed Data Management) system is responsible for the management and distribution of data across the different grid sites. The data is generated at CERN and has to be made available as fast as possible in a large number of centers for production purposes, and later in many other sites for end user analysis. Monitoring their data transfer activity and availability is an essential task for both site administrators and end users doing analysis in their local centers.

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.

Focusing mostly on data management on the grid, the most relevant services for this area of the dashboard are the transfer services and storage managers. It is essential that all information can be easily and quickly propagated to the dashboard service, either directly or via the DDM services, so that end users can have an almost real-time view over their activities and production systems can rely on the system views provided by the monitoring.

File transfer information is transient in most cases, and taken from the main transfer tool being used - the File Transfer Service (FTS). Storage and storage space information lies in the Storage Resource Managers (SRM), which should be able to provide a unique and implementation independent over the physical data and available space. Information regarding file and system meta data is expected to be kept consistent everywhere, and any changes to be propagated to the interested services - like the dashboard.

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 main problem dealing with grid data management tools today is tracking the source of errors. It is very complicated to understand the cause of a file transfer failure, even to identify the service or service class that is responsible for the error, or to distinguish a service problem from a user mistake. We expect the FTS and the new SRM interface to be able to provide and expose to end users better and more consistent error categories, essential to reduce the effort needed today.

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

Data management using the grid depends on a complex set of services. File catalogs for file and file location bookkeeping, transfer services for file movement, storage managers and others. In addition there are several flavors of each of these components, tens of sites each managing a distinct installation - over 100 at the present time - and in some organizations data is seen and moved in larger granularity than files - usually called datasets, which makes the successful usage of the standard grid monitoring tools a non straightforward task.

The dashboard provides a unified view of the whole data management infrastructure, relying mostly on the Atlas data management (DDM) system to collect the relevant information regarding dataset and file movement between the different sites, but also retrieving information from the grid fabric services where appropriate. This last point makes it an interesting tool also for other communities that rely on the same lower level grid services.

Authors: GAIDIOZ, Benjamin (CERN); ANDREEVA, Julia (CERN); SAIZ, Pablo (CERN); ROCHA, Ricardo (CERN)

Presenter: ROCHA, Ricardo (CERN)

Session Classification: Data Management

Track Classification: Data Management