

User-Oriented Data Management in the Scientific Portal Building on the PACI and Teragrid Foundation

Extended Abstract

Alliance All Hands 2003 Poster Session

April 30, 2003

***Jay Alameda, Dennis Gannon, Shawn Hampton, Beth Plale,
Al Rossi, Bob Wilhelmson***

It has become evident through our work with Scientific Portals over the past couple of years that the data management problem could be better addressed with a solution that leverages client-side proximity to the user as a supplement to existing server-side support. In particular we envision a solution where client-side and server-side solutions work together seamlessly. To illustrate the unique position the portal is in with respect to data management, one need only reflect on the volume of information that passes through a portal to and from users and server-side services. Server-side data management services already exist and are being ably provided by the likes of the PDQ expedition and the Teragrid data management group. We suggest augmenting that work with client-side support that leverages server side metadata management systems whenever possible.

The purpose of this poster is to introduce data-oriented 'session management' support for scientific portal users. Intuitively, a session is a duration over which takes place work directed toward a particular well-defined end. A session might be a set of experiments leading up to a paper submission. Other work that the scientist engages in during that period of time, such as preparation for class, would not be part of the session. The session support we envision provides a user with a session-oriented view of his/her data and metadata.

To achieve data oriented session management, we propose two client-side components: a myGridContext and a Teragrid Data Service. MyGridContext is a data management tool that captures data flowing through the portal, manages data and metadata, and assists the user in the recording of experiment results. The second component, Teragrid Data Service, understands the protocols spoken by the various data services (*i.e.*, NeesGrid, MCS, SRB, and gridFTP) and, relying on MyGridContext for contextual information, resolves the location of files and effects their move at the scheduled time.

This whitepaper, which augments an All Hands 2003 meeting poster, illustrates the functionality of the two components by use of an example from scientific computing. Suppose a scientist wants to rerun an experiment with new starting parameters. From the portal she pulls up an existing startup script that contains parameters used in a prior run.

She adjusts particular parameters, and then submits the script through the Portal to a broker service. MyGridContext in the above phase is responsible for such tasks as retrieving the job description, and creating a new session or associating the script with an existing session.

The Broker Service, not part of this effort, communicates a simply stated query *need file 'x' put to location 'l' by time 'y:00'* to the Teragrid data service. 'X' is completely location transparent; it is up to Teragrid Data Service to resolve 'x'. 'l' and 'Y:00' are assumed to have been resolved by the broker service. Teragrid Data Service invokes myGridContext for help in resolving 'x'. Context could be something as simple as a record noting that all input files for this experiment reside in MCS. Teragrid is responsible for contacting the proper metacatalog, and moving the file to its location at the right time. If scheduled moves are provided as part of a metacatalog service, Teragrid Data Service will use those facilities.

Resulting products from the run must be described in detail. Some of that metadata can be captured from the session information we maintain, from monitoring portal activity, and from user interactions through the portal. Any additional information must be asked of the user. The key in collecting metadata is to store it in such a way that the results are highly amenable to later augmentation of additional detail, and highly amenable to efficient searches on a broad set of criteria. The resulting data sets themselves, and the bulk of the metadata we anticipate will be stored at a server side metacatalog.

Processes that generate derived products may either be triggered by workflow or invoked through the session. The session captures and records the relationship between the derived product and its source. Obviously workflow must cooperate with the session manager in order to capture the same information.

A strength of myGridContext is that it provides a rich query interface and single point of contact, the combination of which provides a means by which downstream processing can retrieve the right set of files for processing. For instance, the first step to data mining is in retrieving the right set of files.

MyGridContext and Teragrid Data Service provide the missing puzzle piece between server side data management services and a scientist's interactions with his/her application through the portal. The key concepts are maintaining and recording a user's session in such a way that enables rich and efficient querying of data, and of leveraging the rich support of metadata services such as MCS and NeesGrid metacatalog service whenever possible so as to leverage and expand upon the strengths of these products.

For additional information, please contact Jay Alameda (jalameda@ncsa.uiuc.edu), Dennis Gannon (gannon@cs.indiana.edu), or Beth Plale (plale@cs.indiana.edu).