Geophysical Research Abstracts,
Vol. 10, EGU2008-A-06724, 2008
SRef-ID: 1607-7962/gra/EGU2008-A-06724
EGU General Assembly 2008

Flexible Climate Data Analysis with C3Grid

U. Ulbrich (1), I. Kirchner (1), H. Kupfer (1), A. Papaspyrou (2), C. Grimme (2) and the C3Grid-Team

(1) Freie Universität Berlin, Institute for Meteorology (uwe.ulbrich@met.fu-berlin.de), (2) Technische Universität Dortmund, Robotics Research Institute, Section Information Technology

In climate research, there is a growing focus on the evaluation of multi-model ensemble simulations, which at the same time comprise of data with higher spatial and temporal resolution, meaning strongly growing data amounts. Before these data can be processed, a regular user has to identify the location of their storage systems as well as the formal and logical data formats. In most cases, processing of the data cannot be done at the storage system itself, so that transfer of the data to a suitable processing location is necessary. These steps are part of the regular workflow of data analysis which is time consuming and hampers the scientific process. The C3Grid (Collaborative Climate Community Data and Processing Grid) has been designed to alleviate this problem. It permits an identification of datasets and their location through a portal, making use of standardized metadata, and provides tools to extract and process the data needed. Here, we focus on the flexibility introduced into the C3Grid in order to permit the incorporation of a wide range of individual workflows that are built from a number of consecutive “atomic” tasks, and a parallel processing of such tasks. It is shown how the problem of complex dependencies between those tasks is solved automatically, handling the workflow execution with a Workflow Scheduling Service (WSS). In the context of C3Grid a workflow is defined as a set of atomic tasks with dependencies between them and it is described by using the C3Grid-specific XML dialect. The set of contained atomic tasks can be divided in three classes: data extraction, transfer and execution tasks. Those are defined with the help of the Open Grid Forum-standardized Job Submission Description Language (JSDL) along with Earth Science specific extensions. Such a workflow description is given as input to

the WSS which analyzes and validates it and handles the execution of the contained tasks according to the defined dependencies. In addition to this basic functionality, the WSS optimizes the allocation of computing and data extraction resources as well as the automatic handling of intermediary data transfers. That is, the process of mapping computing tasks to specific resources is supported by a Job Scheduler (JS) component and implicates the autonomous just-in-time creation of intermediary transfer tasks by utilizing the workflow dependencies. The information on available resources which the WSS needs for performing the optimization is collected via a central Grid information service. Subsequently, data handling and transfer is executed by the external and independent Data Management Service (DMS).

The before described functionality is reflected in the loosely coupled and service-oriented architecture of the WSS. Each workflow submitted to the C3Grid is represented by a Job Management Service (JMS) instance which handles the analysis, validation, and internal representation of the submitted workflow. Each atomic task out of a workflow is then assigned to an individual Task Execution Service (TES) instance that handles all task related activities: transportation of data, execution of analysis or specific data extraction. The decision where and when to start a task execution is made by the scheduling component and its individual strategy. By this design a distributed, parallel, and completely decoupled execution of workflows is enabled.

The implementation of these services is based on the Globus Toolkit 4.x Framework for stateful web services (WSRF). This basis ensures compliance to widely used standards in service definition and communication. Further, it simplifies the monitoring of a workflow status by supporting the WS-Notification standard. Herewith, a user does not have to permanently supervise the processing of the workflow but is notified by the WSS about completion or failure. Possible ways of communication include messages via email or mobile communication devices.

The C3Grid is thus particularly designed to enable user submission of own workflows which are compiled from standard tasks. The possibilities of such an implementation are described using the example of standard GCM data analysis workflows and of a data extraction and compilation workflow for chemical weather forecasts.