ABSTRACT

Many sophisticated environments allow creating and managing of scientific workflows, whereas the workflow itself is provided as a service. Scientific Grids handle large amounts of data and share resources, but the implementation of service-based applications that use scientific infrastructures remains a challenging task, due to the heterogeneity of Grid middleware and different programming models. This paper proposes an e-Science environment providing functionality in a simplified way, considering the Grid as a source of computational power and an information infrastructure. To promote integration among components and user interaction, the paper outlines a SOA-based scientific environment where an experiment is modeled through an abstract workflow defining the functional model of the experiment. The tasks are mapped to the corresponding scientific services by a workflow engine, separating logical aspects from implementation issues. Services depend on the type of experiment and can be re-used, wrapped, or moved into a new workflow. Infrastructural services discover suitable resources that match user requirements and schedule workflow tasks. Further, they monitor the execution of each task and aggregate the results. The proposed approach provides a simple-to-use and standardized way for the deployment of scientific workflows in a distributed scientific environment, including the Grid.
INTRODUCTION

Collaboration in scientific experiments, obtained by sharing data, tools, and expertise towards a common scientific goal, is becoming more and more appealing for e-Science, thanks to the availability of Information and Communication Technology (ICT) methods and tools. In particular, the Service Oriented Architecture (SOA) paradigm is attractive since it can effectively support distributed cooperation. However, the heterogeneity and dynamicity of services and of their underlying infrastructures make the aspects of creating valuable complex service environments an emerging research issue in the scientific community.

Advances in computing technologies have enabled scientists to validate new research practices in many scientific fields and to evolve from individual activities to work conducted in teams, exploring research issues at time and space scales both greater and finer than ever before. This new research context is becoming more and more complex in terms of the number of collaborating researchers, the diversity of computing environments supporting collaborative efforts among each participant in data/computation intensive applications, the number of emerging powerful and effective data analysis tools enabled by new technologies, the distribution of data and computing resources and the consequent orchestration of the data analysis tools across various platforms.

Indeed, the computing resources available to a scientific experiment, the network capacity, connectivity and costs may all change over time and space since some components are added, removed or temporary unavailable. Similarly, the scientist may move from one location to another, joining and leaving groups of researchers and frequently interacting with computers in changing experimental situations. In short, the research environment we consider is constantly in evolution and scientific collaboration keeps on increasing the aggregation and sharing of heterogeneous and geographically dispersed resources. In practice, this means that computation does not occur at a single location in a single context, but rather spans a multitude of situations and locations covering a significant number of heterogeneous hardware or software components.

E-Science is the term usually applied to the use of advanced computing technologies to support scientists. In short we can say “e-Science is about global collaboration in key areas of science, and the next generation of infrastructure that will enable it” (De Roure, 2004). The above definition is still to come at the structural level: technical problems limit the usability of the e-infrastructure presently in production, i.e., the Grid, whose technology is still far from allowing a true interoperability of scientific applications and/or computational experiments. As a consequence, the level of detail needed for the successful deployment of scientific applications on the Grid still remains very high. Moreover, scientists want to get work done and they do not want to deal with the complexity of building applications that expose details of the underlying e-infrastructure. They must be able to express their problem by composing application specific components in an easy-to-use, easy-to-re-use and easy-to-modify form. Their favorite model of programming is to compose a workflow by means of a graphical interface via “drag-and-drop”, and they loathe writing “programs” in XML. However, the visual programming model must be sufficiently powerful to address a wide range of conditions, exceptions, iteration and adaptive control.

The paper aims at defining the needs and the building blocks for the next step in the advance of e-Science environments. Grids and distributed systems, augmented with various management capabilities, are considered essential aspects of the e-Science environment. To promote both integration among components and user interaction, the paper proposes to extend the use of solutions developed for business environments and in particular the adoption of a Service Oriented Architecture. An architectural model for the de-