ABSTRACT

The increasing ability for the sciences to sense the world around us is resulting in a growing need for data-driven e-Science applications that are under the control of workflows composed of services on the Grid. The focus of our work is on provenance collection for these workflows that are necessary to validate the workflow and to determine quality of generated data products. The challenge we address is to record uniform and usable provenance metadata that meets the domain needs while minimizing the modification burden on the service authors and the performance overhead on the workflow engine and the services. The framework is based on generating discrete provenance activities during the lifecycle of a workflow execution that can be aggregated to form complex data and process provenance graphs that can span across workflows. The implementation uses a loosely coupled publish-subscribe architecture for propagating these activities, and the capabilities of the system satisfy the needs of detailed provenance collection. A performance evaluation of a prototype finds a minimal performance overhead (in the range of 1% for an eight-service workflow using 271 data products).
INTRODUCTION

The need to access and share large-scale computational and data resources to support dynamic computational science and agile enterprises is driving the growth of Grids (Foster, Kesselman, Nick & Tuecke, 2002). In the realm of e-Science, *science gateways* are archetypes for accessing, managing, and sharing virtualized resources to solve large collaborative challenges (Catlett, 2002; Gannon et al., 2005). Science gateways are built as a *service-oriented architecture* with Grid resources virtualized as services. These resources—including physical resources such as sensors, computational clusters, and mass storage devices, and software resources such as scientific tasks and models—are available as services that provide an abstraction to access the resources through well-defined interfaces.

A significant constituent of applications that make use of the science gateways is *data-driven applications* (Simmhan, Pallickara, Vijayakumar & Plale, 2006a). The proliferation of wireless networking and inexpensive sensor technology is allowing the sciences an increasing ability to sense the world around us (West, 2005). This is specifically resulting in a growing need for data-driven applications; that is, applications that can be computation-intense and are usually either dataflow applications in which data flows from one process to another, or demand-driven in which computations are triggered in response to events occurring in the world around us. Data-driven scientific experiments are designed as workflows composed of services on the Grid, and data flow from one service to another, being transformed, filtered, fused, and used in complex models. These workflows capture the invocation logic for the scientific investigation and may be composed of hundreds of services connected as complex graphs. Data-driven workflow executions also see the participation of thousands of data products that reach terabytes in size. At this scale of processing, users need the ability to automatically track the execution of their experiments and the multitude of data products created and consumed by the services in the workflow. Provenance collection and management, also called process mining, workflow tracing, or lineage collection, is a new line of research on the execution of workflows, and the derivation and usage trail of data products that are involved in the workflows (Bose & Frew, 2005; Moreau & Ludascher, 2007; Simmhan, Plale & Gannon, 2005).

*Provenance* collected about the tasks of a workflow describes the workflow’s service invocations during its execution (Simmhan et al., 2005). This helps track service and resource usage patterns, and forms metadata for service and workflow discovery. In data-driven applications, however, it is provenance about the data that is central to understanding and recreating earlier runs. In *data-driven workflows*, data products are first-class parameters to services that consume and transform the input data to generate derived data products. These derived data products are ingested by other services in the same or a different workflow, forming a data derivation and data usage trail. *Data provenance* provides this derivation history of data that includes information about services and input data that contributed to the creation of a data product. This kind of information is extremely valuable, not only for diagnosing problems and understanding performance of a particular workflow run, but also to determine the origin and quality of a particular piece of derived information (Goble, 2002; Simmhan, Plale & Gannon, 2006b).

Current methods of collecting provenance are from workflow engine logs (IBM, 2005) or by instrumenting the services (Bose & Frew, 2004; Zhao, Wroe, Goble, Stevens, Quan & Greenwood, 2004). In the former case, the logs from the workflow engine are at the message level and insufficient for deciphering provenance about the data products, while instrumenting services introduce a burden on the service author to modify their service to generate provenance metadata. They also tend to be specific to the workflow framework and are not interoperable with heterogeneous workflow models that are likely