Provenance in Kepler-based Scientific Workflow Systems

Ilkay Altintas[4], George Chin[5], Daniel Crawl[4], Terence Critchlow[5], David Koop[2], Jeff Ligon[1], Bertram Ludaescher[3], Pierre Mouallem[1], Meiyappan Nagappan[1], Norbert Podhorszki[3], Claudio Silva[2], Mladen Vouk[1]

Introduction

- Scientific workflow management systems are used to automate the data management and analysis tasks of scientific discovery.
- Increasing complexity of such workflows, and sometimes legal reasons, is fueling a demand for more run-time and historical information about the workflow processes, outputs, environments, etc.
- Properly constructed run-time and provenance information collection framework can help manage, integrate and display the needed information.
- In this poster we present the current provenance system developed by the Department of Energy Scientific Data Management Enabling Technology Center's Scientific Process Automation group.

Definition of Terms

- *Data Provenance* focuses on data flows, data history, inputs, outputs, and data transformations that occur during scientific workflow execution
- *Process Statistics* show how *progress* is made through the workflow control flows and event flows (sequence diagrams) occurring during workflow execution
- *Workflow Evolution* is about the different versions and implementations of the workflow, i.e. about evolution of its structure and form
- *System Information* is data about system environments in which a workflow executes

System Architecture

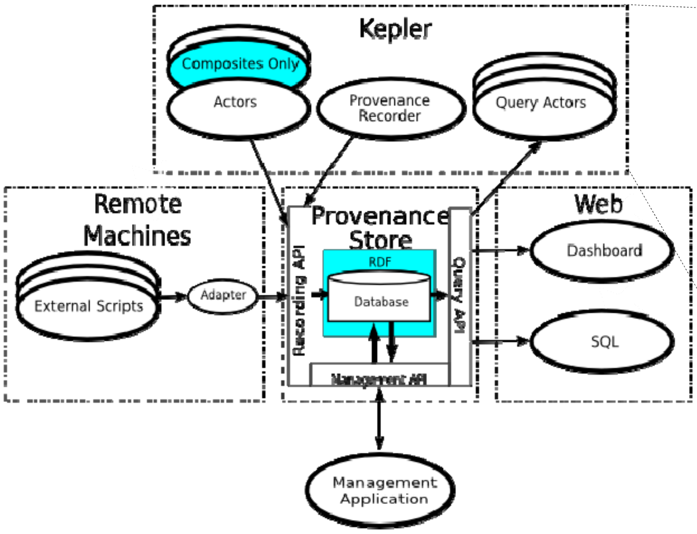
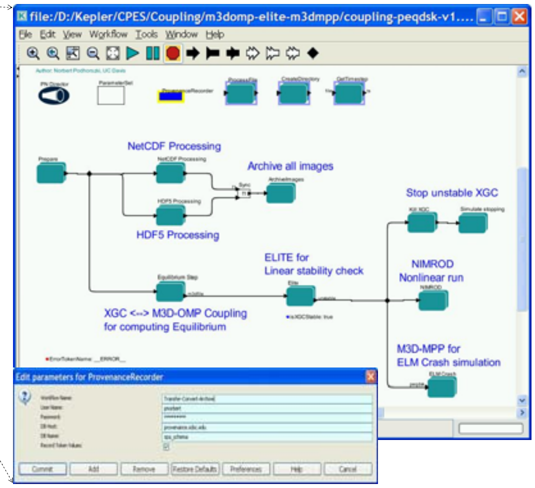


Figure 1 – System Architecture



Workflow Name: Transfer-Convert-Archive
User Name: pnorbert
Password: *********
DB Host: provenance.sdsc.edu
DB Name: sps_schema
Record Token Values: ☑

Figure 2 – Provenance enabled Kepler

Challenges & Solutions

- How to collect provenance information in a standardized and seamless way and with minimal overhead – Modularized design and integrated provenance recording
- How to store this information in a permanent way so that the scientist can come back to it at anytime, - Standardized schema
- How to present this information to the user in a logical manner – an intuitive user web interface: Dashboard
- How to implement security policies that apply to Department of Energy (DoE) national laboratories – One time passwords, pushout communications and encapsulated resources.

Architecture

- The solution adds to the successful Kepler scientific workflow support system by integrating Kepler with a standard *LAMP - Linux Apache MySql PHP* environment to provide a very flexible and readily deployable (K)LAMP scientific workflow support environment for e-science.
- The solution is sufficiently modular to allow use of other workflow engines and other component solutions.

Uses

- Data Auditing, Debugging of a workflow, Steering the execution of the workflow, Crash Recovery, etc.

This work sponsored in part by NC State DURP funds, the DOE SciDAC grant DE-FC02-01ER25484, and the IBM Corp. Shared University Research Program.

Universities: 1. NCSU, 2. UUtah, 3.UC Davis
Labs: 4.SDSC, 5.PNNL

19-Oct-07/V5