Supporting Tool for Collaborative Scientific Workflow

D.Khadar Hussain[1] and S.Md.Haroon[2],
[1]M.Tech, Research Scholar, CSE Dept and [2]M.Tech, PG Scholar,
[1]JNTUA College of Engineering,
Anantapuramu-515002.A.P.

Abstract: Collaboration has become a dominant feature of modern science. Many technical problems are outside the realm of individual discipline or scientist to solve and hence require cooperativedeterminations. Recenttechnical data management and investigation usually rely on multiple scientists with variouscapabilities. In current years, such a cooperativedetermination is often structured and automated by a dataflow-oriented process called technical workflow. Existing tools are single user-oriented and do not support workflow development in a collaborative fashion. Based on scientific collaboration ontology, we suggest a service-oriented collaboration model supported by a set of compos able collaboration primitives and designs. The collaboration procedures are then applied to support effective concurrency control in the process of collaborative workflow composition. The Project and growth of Confucius a service-oriented collaborative scientific workflow composition tool that extends an open-source, single-user environs.

Index Terms— H.4.1.g Workflow organization,M.4 SOA, Computer-supported collaborative work, Confucius Tool.

I. INTRODUCTION

Recenttechnical data management and investigation usually rely on multiple scientists with differentknowledge. In current years, such a cooperative effort is often composed as and automated by a dataflow-oriented process called systematic workflow. Currentworktables are single user-oriented and do not support scientific workflow application development in a "collaborative fashion."

The advancement of modern science has created sheer volume of data with increasingcomplexity. Processing and managing such largescale scientific data sets is usually beyond the realms of individual scientists to solve; instead, it has to rely on multiple domain scientists with diverse expertise. For example, the Large Synoptic Survey Telescope (LSST) experiment, which aims to repeatedly image half of the sky over a planned 10-year survey, produces data at a rate of 300 MB/s and will result in catalogs of about 130 TB of roughly 3×109 sources times 10 years worth of data. Analyzing such data sets demands a collaboration of a number of organizations with over 1,800 scientists and engineers engaged. Such scientific data analysis and processing is usually structured and automated by a dataflow-oriented process called scientific workflow. In contrast to business processes that are control-flow oriented andorchestrate a collection of well-defined business tasks to achieve a business goal, scientific workflows are often dataflow-oriented and streamline a collection of scientific tasks to enable and accelerate scientific discovery. Researchers use scientific workflows to integrate and structure local and remote heterogeneous computational and data resources to perform in silico experiments.

Fig.1: Work flow Model

Collaborative workflow design is usuallycritical for the success of a comprehensiveworkflow composition. A very simplified LSSTexperiment represents a three-step workflow: dataretrieval, pre-processing, and data modeling.

Designing each step requires different expertise.Meanwhile, the serial relationship among the stepsimplies that the accuracy of the entire process isdetermined by the design of each step and the errorpropagation between them. While involvingscientists with different capabilities focus onfinding a local optimal design of a particular step,collaboration between them will find a globaloptimal design for the entire workflow. We focuson reporting our design and development of onekey enabling technique, collaboration protocol. Wepropose scientific collaboration provenanceontology, and based on it, we have developed aservice-oriented collaboration model that issupported by a set of compos able collaborationprimitives and designs. The teamworkproceduresare to support effective concurrencycontrol in the process of workflow co-design. Thecore idea of Service Oriented Architecture (SOA) isto position services as the primary means(components) that encapsulate solution logic asreusable assets. In this project, we have leveragedthe concept of SOA to build our system for higherinteroperability, reusability, and productivity. Sincewe focus on scientific workflows, throughout thispaper, we use the terms scientific workflows andworkflows interchangeably.

2. RELATED WORK

2.1 Management workflow system

To date, several scientific workflowmanagement systems (SWFMSs) have beendeveloped as single-user environments that helpindividual scientists construct workflows fromavailable scientific resources. RepresentativeSWFMSs include Kepler, Taverna, Triana,VisTrails, Pegasus, Swift, Trident, and VIEW, Eachsystem shows singlestructures. It provides a platform to support individualscientists in composing. Some systems show somecollaboration features, in the sense that they allow ascientist to compose a workflow from sharedresources and services. However, they providelimited support for multiple scientists tocollaboratively compose a shared workflow.

2.2 Coordination among Business Process

For business workflows, the term "collaborative workflows" is interchangeable withthe term "coordinated workflows" .They emphasize coordination between workflows. The Computer Supported Cooperative Work (CSCW) community has studied the general-purpose concurrent enterpriseproblematic, where coworkers with differenttitles possess different controls over the shared work product. To name a few, OntoEdit supports a collaborative software engineering process; Yen et al. present a collaborative design tool that allows privileged collaborators to change the process; OPCA Team integrates the objectoriented and process-oriented paradigms into onesingle framework to enable the coexistence of structured processes and human interaction behaviors in one business process modeling system. In contrast, our work focuses on dataflow-oriented collaboration, where semantic relationships and constraints between different comprising components (e.g., tasks and data links) need to be carefully considered during concurrent composition.

3. COOPERATIVE SCIENTIFIC WORKFLOW COMPOSITION MODEL

Provenance has been widely considered critical to the reproducibility of scientificworkflows. Compared to existing significant amount of work focusing on provenance for runtime workflow execution, our work focuses on collaboration provenance that tracks human interactions and efforts in the process of scientific workflow composition. Our method is to record all collaborative activities that contribute to a composed workflow.

3.1 Cooperative Composition Model

We designed models to regulate how collaborators can collaboratively design and update mutual workflows. Instead of reinventing the wheel, we chose to explore how to extend the single user-oriented Taverna tool.

3.2 Advanced collaboration model

Scientific collaborations usually last for a long period of time, e.g., months and years. In addition, temporary discussion groups and sessions may be formed in the lifecycle of a long-term scientific collaboration process.

Fig.2: Cooperative workflow composition Provenance ontology

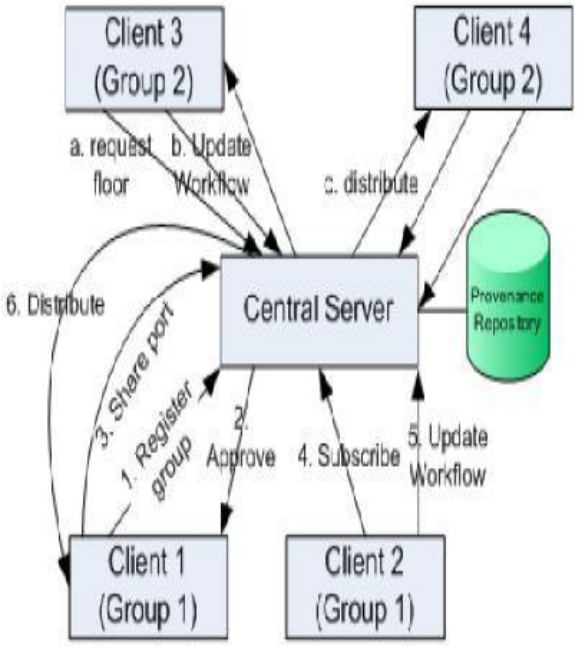

Fig.3: Collaborative composition model

Therefore, we constructed a hierarchical structure for the central server. It may host multiple collaboration groups, which may or may not have nesting relationships between them. The central server maintains all collaboration group information and acts as the subject for all registered groups. All observers (collaborators) are organized into corresponding collaboration groups. The central server also stores and manages all provenance data, so that it becomes a repository of workflow products and enablesscalability. In other words, we realize a multitenancy infrastructure.

4. A DATAFLOW-BASED SCIENTIFIC WORKFLOW COMPOSITION MODEL

Our proposed locking scheme is based on our previous dataflow-based scientific workflow composition model. Here, we consider workflow-level composition, in which workflows are the only operands for workflow composition. A set of workflow constructs including unary constructs have been proposed such as Map, Reduce, Tree, Conditional, Loop, and Curry. These constructs transform workflows to new workflows.

A workflow definition includes a workflow interface and a workflow body. A workflow interface declares the workflow identifier, workflow name, description, and input/output ports. A workflow body defines the implementation of the workflow. There are currently three kinds of implementations: primitive workflow, unary-construct based workflow, and graphbased workflow body. Primitive workflows are the basic building blocks of our model which are constructed from tasks. Unary-construct based workflows are created by applying unary constructs on existing workflows. Graph basedworkflows are defined from a workflow graph which areconstructed from a set of workflows and a set of data link constructs.

5. SCIENTIFIC WORKFLOW LOCKING SCHEME

A scientific workflow *w depends* on scientific workflow *w_*, denoted as $w\_ \leftarrow w$, if they satisfy one of the following conditions:

• There exists a construct $c{:}W \rightarrow w$ and $w\_ \in W$. Then *w* is a *parent* of *w_*, and *w_* is a *child* of *w*.

• There exists a sequence of workflows *w*1, *w*2... *wk*such that *w*_ ← *w*1, *w*1 ← *w*2... *wk*← *w*. Then *w* isan *ancestor* of *w*_.

A scientific workflow is a *composite workflow* if it has at least one child and a workflow is a *primitive workflow* if it has no families. A technical workflow is a *root workflow* if ithas no parents.

A scientific workflow dependency graph is a finite set *W* of nodes and a finite set *D* of edges (a subset of *W* × *W*), such that each node represents a workflow and each edge represents a dependency relationship. A scientific workflow dependency graph is *well-formed* if it containsexactly one root node. A valid scientific workflow composition must contain one and only one root workflow because there must exist a main workflow as the entry of the composition. Therefore *ascientific workflow dependency graph generated by a validscientific workflow composition is well-formed*.

6. IMPLEMENTATION OF SYSTEM

We have constructed a collaboration pattern template library. The basic building blocksare collaboration primitives. Users can build new collaboration patterns using existing collaboration primitives. Identified collaboration patterns are stored as provenance data to support the tracking, storing, and querying of interactions and coordination among scientists. Without reinventing the wheel, we extended the single user-version Taverna into a collaborative version. The reason why we chose Taverna is mainly based on itspopularity and large user base. Another reason is that Taverna is an open-source tool developed in Java. Thus we can explore its code base. We built a central server supporting all workflow collaborations. Workflow evolution provenance and collaboration provenance are storedin a shared database on the server. Eachcollaborator may store an intermediate version ofthe workflow at a local machine, but all committedactivities are stored at the server, to supportasynchronous collaboration where collaboratorsmay work on the shared workflow at preferable time. We consider four options for selecting database systems: native XML, relational, XMLrelational, and RDF. Currently we use a relational database because it is a preferable choice of Taverna, upon which our Confucius is built.

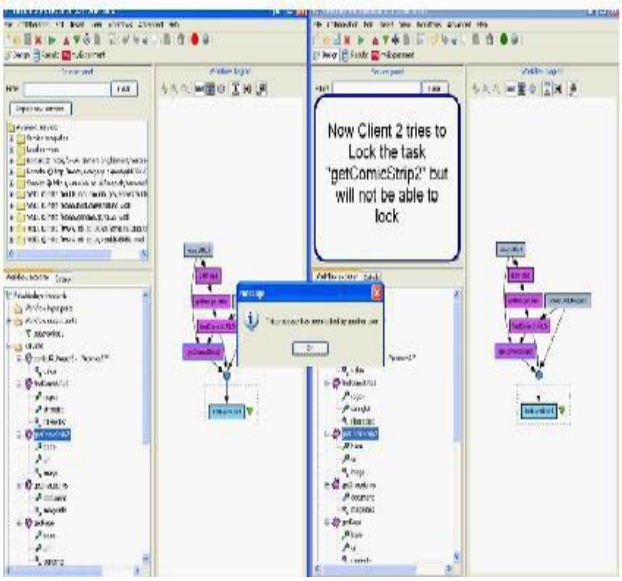

Fig.5: Screen shots of concurrent workflow updates

7 .CONCLUSION

With the advent of ongoing work on establishing collaboration protocols to supportcollaborative scientific workflow composition, Our service-oriented infrastructure includes a collaboration ontology associated with a set of collaboration patterns, primitives, and constructs, as well as concurrent control mechanisms to supportconcurrent collaborative workflow composition. We plan to continue our research in the following directions. First, we will design and conduct an evaluation study and use the feedback to enhance the system. Second, based on the collaboration we plan to enhance collaboration provenance management performance. Third, weplan to conduct more experiments to study the effects of tuning various parameters (e.g., the number of concurrent collaborators, the productivity of individual members, the number of tasks comprised in the shared scientific workflow) on concurrent efficiency. Fourth, we plan to discovershowingcooperative scientific workflow composition in the Cloud infrastructure.

We then proposed a formal granular scientific workflow locking scheme and algorithms for locking and releasing workflows and constructs ourproposed scheme guarantees the correctness of concurrentworkflow update, in the future, we plan to propose techniquesto further validate the consistency of the resulted workflow compositions.We also plan to

study runtime issues of collaborative workflow orchestration and coordination.

REFERENCES

[1] S. Wuchty, B. Jones, and B. Uzzi, "TheIncreasing Dominance of Teams in Production ofKnowledge", *Science*, 2007, 316: pp. 1036- 1039

[2] LSST, "Large Synoptic Survey Telescope",2009, Accessed on, Available from: http://www.lsst.org/lsst/science.

[3] Y. Gil, E. Deelman, J. Blythe, C. Kesselman,and H. Tangmunarunkit, "Artificial Intelligence andGrids: Workflow Planning and Beyond", *IEEEIntelligent Systems*, Jan.-Feb., 2004, 19(1): pp. 26–33.

[4] E. Deelman and Y. Gil, "NSF Workshop on theChallenges of Scientific Workflows", (ed.), May 1-2, 2006.

[5] B. Ludascher, "Scientific Workflows:Cyberinfrastructure for e- Science", in Proceedingsof *Pacific Neighborhood Consortium (PNC)*, 2007,Berkeley, CA, USA, Oct. 19, pp.

[6] G.M. Olson, A. Zimmerman, and N. Bos, *eds.,Scientific Collaboration on the Internet*, 2008, MITPress, Cambridge, MA, USA.