INTERNATIONAL JOURNAL OF MERGING TECHNOLOGY AND ADVANCED RESEARCH IN COMPUTING

ISSN: 2320-1363

Supporting Collaborative Tool of A New Scientific Workflow Composition

Md.Jameel Ur Rahman*1, Akheel Mohammed*2, Dr. Vasumathi*3

Abstract

Large scale scientific data management and analysis usually relies on many distributed scientists with diverse expertise. In recent years such a collaborative effort is often composed and automated into a dataflow oriented process so called scientific workflow. Collaboration has become a dominant feature of modern science. Many scientific problems are beyond the realm of individual discipline or scientist to solve and hence require collaborative efforts. Mean while today's science becomes increasingly more data-intensive resulting in rapid transition from computational science to e-science. However the existing scientific workflow tools are single user oriented and do not support collaborative scientific workflow composition, execution, and management among multiple distributed scientists. We propose a collaboration model supported by a set of collaboration primitives and patterns. The collaborative protocols are then applied to support effective concurrency control in the process of collaborative workflow composition. This article presents a disciplinary definition of this term, discusses the opportunities, requirements, and challenges of collaborative scientific workflow for the enablement of scientific collaboration and concludes with our ongoing work in this direction.

Keywords: **Scientific Collaboration, Scientific Workflow, Collaborative Scientific Workflow.**

1. Introduction

In recent years scientific have started to use scientific workflow to integrate and structure local and remote heterogeneous computational and data resources to perform in silico experiments and have made significant scientific discovers. I contrast business workflows that are control flow oriented and orchestra a collection of well-defined business tasks to achieve a business goal, scientific workflow are dataflow oriented and streamline a collection of scientific tasks to enable and accelerate scientific discovery. Several scientific workflow management systems (SWFM) have been developed to support scientific workflow design and execution such as Kepler, Taverna. Existing SWFMs mainly support single scientists to compose and execute scientific workflow. Modern scientific research projects, however, are collaborative in nature, and team members usually reside at geographically distributed locations. Meanwhile such extreme scale scientific data analysis and processing is usually composed and automated into a dataflow oriented process so called a scientific workflow. Scientists use scientific workflow to integrate and structure local and remote heterogeneous computational and data resources to perform in silico experiments. To facilitate more interactivity between collaborators to better support exploratory collaborative data analysis and enable effective steering of the computational process in the context of scientific workflows we have been developing a collaborative scientific workflow tool. In this article we present the preliminary result of our study of collaboration protocols supporting effective and efficient collaborative scientific workflow composition. We propose scientific collaboration provenance ontology and base on it, a collaboration model supported by a set of collaboration primitives and patterns. The collaboration protocols are then applied to support effective concurrency control in the process of collaborative workflow composition.

2. Related Work

We compare out approach with related work in three categories: scientific workflow management system, business workflow coordination, and collaborative workflow composition. The business community recently recognized the need of involving humans into business workflow and has developed a preliminary model. However the model is inapplicable to collaborative scientific workflow due to the fundamental difference between business workflow and scientific workflows. While the business workflows are control flow oriented scientific workflow are dataflow oriented. Furthermore provenance data management for the reproducibility of scientific results is essential for scientific workflows but not for business workflows. Hence scientific workflows pose a different set of requirements. With the rapid emergence of services computing technology a workflow may select optimal available services at runtime based in some QoS measurements. We conclude that workflow control should be driven it should be customizable and adaptive during runtime.

2.1 Scientific Collaborative Provenance Ontology

We develop scientific collaboration provenance ontology to support the modeling of various traditional provenance information about scientific workflow and user interaction and collaboration patterns. The ontology is shown in fig.1. The latest advance of IT technologies enabled and encouraged people to from large scale and multidisciplinary scientific research projects to solve complex

ISSN: 2320-1363

scientific problems. Establishing a knowledge base our collaboration provenance ontology is centered upon the concept of workflow. Each scientific workflow comprises organized processors and data links as well as predefined requirements and annotations dynamically generated. Each workflow maintains one or more floors that are tokens to ensure concurrently control.

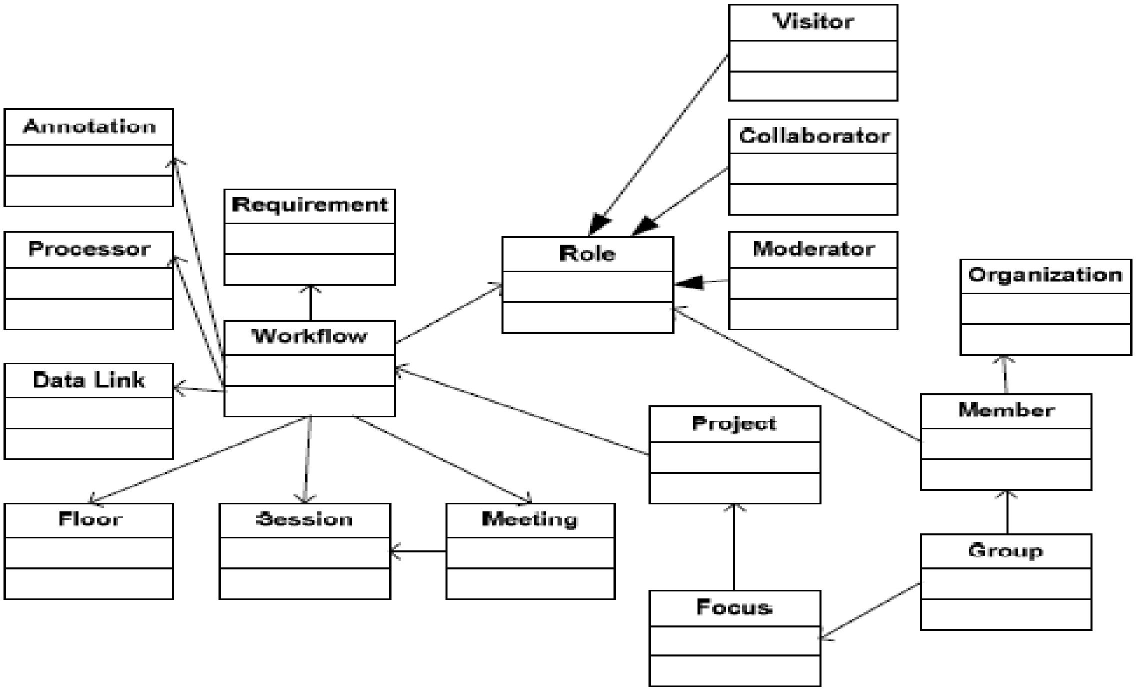


Fig.2.1.1 **A scientific collaboration provenance ontology**.

Long term collaboration on a scientific workflow forms a meeting. A short term synchronous collaboration is called a session. Each scientific workflow belongs to a project. Each project belongs to a scientific group. Each group may comprise multiple focuses, each involving multiple projects. A group contains a set of members each may belong to different organizations. Collaboration on a scientific workflow is conducted by members serving in different roles. The initiator of a scientific workflow is called a moderator. Scientists who cooperate on the lifecycle of a workflow are called collaborators. Our scientific collaborative provenance ontology which is extensible serves as a foundation for managing collaboration provenance.

I. **Collaboration Protocols**
II. **Collaboration Patterns**
III. **Collaboration Model**

Establishing a collaboration is important to support effective human interaction and collaboration throughout the life cycle of scientific workflow composition. Such collaboration model will be independent and can be dynamically plugged into other models to favor configurability and re-configurability. Different scientific research projects may adopt different collaborations rules and patterns.

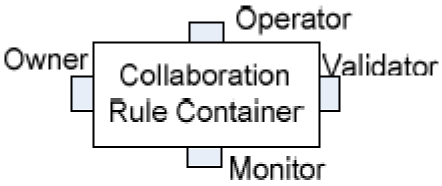


Fig.2.1.2. Collaboration rule container.

C_Rule =<Owner, Operator, Monitor, Validate>

The collaboration container comprises four basics plug in roles: Owner, operator, monitor, and validate. An owner role represents a group of scientists who have ownerships over a dataset or a task.

I. Collaboration Patterns

Significantly different from business collaboration scientific collaboration is typically exploratory and unpredictable thus requiring constant human interaction and intervention in the process. Therefore we have been studying scientific collaboration scenarios to identify data centric collaboration patterns. Dataset request pattern reflects a scenario when some specific data is required during the execution of a scientific workflow to continue the exploration, while the dataset belongs to an external scientific group. These collaboration patterns can be represented using our proposed simple yet powerful collaboration model. For example if applied to a dataset this collaboration container can determine that certain scientists have the ownership over the dataset. Note that our proposed collaboration model shows great flexibility if a scientific reconfigures some parameters of a particular collaboration container ar runtime the collaboration policy affecting the scientific workflow may be changed accordingly.

a) Services Oriented Collaboration Realization

One is to facilitate communication between collaborators. The other is to enable collaboration provenance collection, meaning that the collaboration processes is recorded and can be replayed later on. Thus we construct a uniform collaboration message based communication protocol.

i) Collaboration Primitives

Based on the collaboration patterns we identified a set of semi structured collaboration primitives as summarized. Since the scientific collaboration may last a long period of time we adopt an asynchronous communication mode meaning that each collaboration primitive is associated with an instant acknowledgment. In addition to be used individually these collaboration primitives serve can be used as building blocks for collaborators to model comprehensive collaboration patterns.

ii) Collaboration Mini- Workflow

Based on the identification collaboration primitives, we apply the concept of service oriented architecture to implement the collaboration patterns. Each collaboration pattern is accomplished by a mini workflow comprising a set of configured collaboration primitives. Through combinations over the set of collaboration primitives different collaboration patterns can be realized. Each collaboration primitives is realized by a service call associated with the corresponding messages. Once represented by Business Process Execution Language (BPEL) multiple collaboration constructs may combine to form a comprehensive collaboration scenario. Such a service oriented model enables platform neutral and language neutral collaboration.

b) Service Orientation Collaboration Provenance

Provenance has been widely considered critical to the reproducibility of scientific workflow. Our method is to record all collaborative activities leading to a composed workflow. As the best enable universal communication among participating scientists with platform independence. Messages are divided into requested message and response message. Each message contains one or more primitives that form a transaction, meaning that they from an atomic unit of work in a scientific workflow.

1. Composition Concurrent control

The lifetime of a collaborative scientific workflow may last for a long period of time thus the concurrency control over its different phases deserves consideration.

a) Locking Granularity

The reason why we chose Taverna is mainly based on its popularity and big user base. Another reason is that Taverna is an open-source tool developed in Java. Thus we can explore its code and turns it into a collaborative version. Adopting the instrument from an extensively tested and well proved human communication protocol, Robert's Rules of Order (RRO), we establish a floor control mechanism. Each scientific workflow maintains a single floor (token), which can be assigned to one collaborator at a time. Each collaborator requests and competes for the floor. Only the collaborator holding the floor can propagate their changes on the shared workflow. After done with the update, the collaborator can release the floor and other collaborators may get it. Such a workflow-level floor control may not be efficient to support large-scale scientific workflow composition. Since scientific research is an exploratory process, the development of a scientific workflow may undergo many discussions and changes and may last for a long period of time. Meanwhile, a collaboration group nowadays usually comprises scientists from different organizations at distributed locations. They may possess different schedules and may even reside in different time zones; thus, their collaboration may adopt both synchronous and asynchronous modes. To increase composition concurrency, we investigate the option of locking the smallest building blocks. A scientific workflow allows multiple un-overlapped locks, so that multiple scientists may work on the locked components simultaneously.

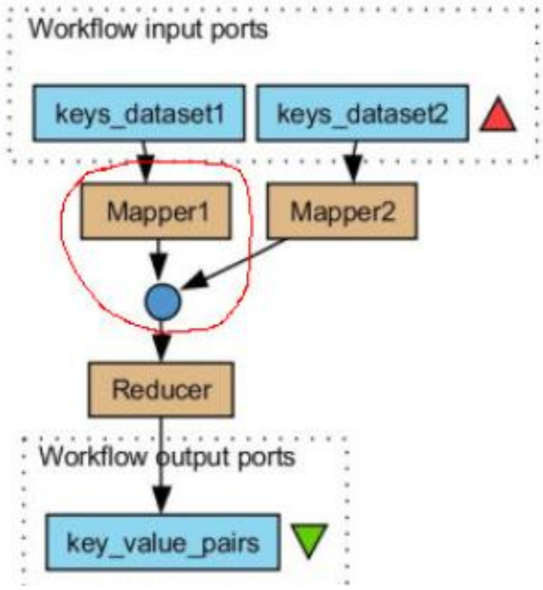


Fig.3. word count workflow.

If we set up the locks on individual processors and data links only, two collaborators may concurrently update one processor Mapper1 and its output data links respectively. This situation may not be desirable, because the data link directly depends on the processor. In other words, connected processors and data links may have close semantic relationships, which need to be preserved by requiring that neighboring entities cannot be updated by different collaborators at the same time.

Therefore, we propose a concept of synchronization area that represents a conceptual area in a shared scientific workflow, which allows only one collaborator to work on it at a given time. Such an area represents a dynamic semantic area. In the context of a Taverna workflow, if a user tries to lock a data link, the synchronization area is the data link. If a user tries to lock a processor, the synchronization area is dynamically delimited and includes all of the fan-out data links of the processor. In Fig. 5, the manually drawn red circle around the Mapper1 processor and its fan-out data link represents such a synchronization area.

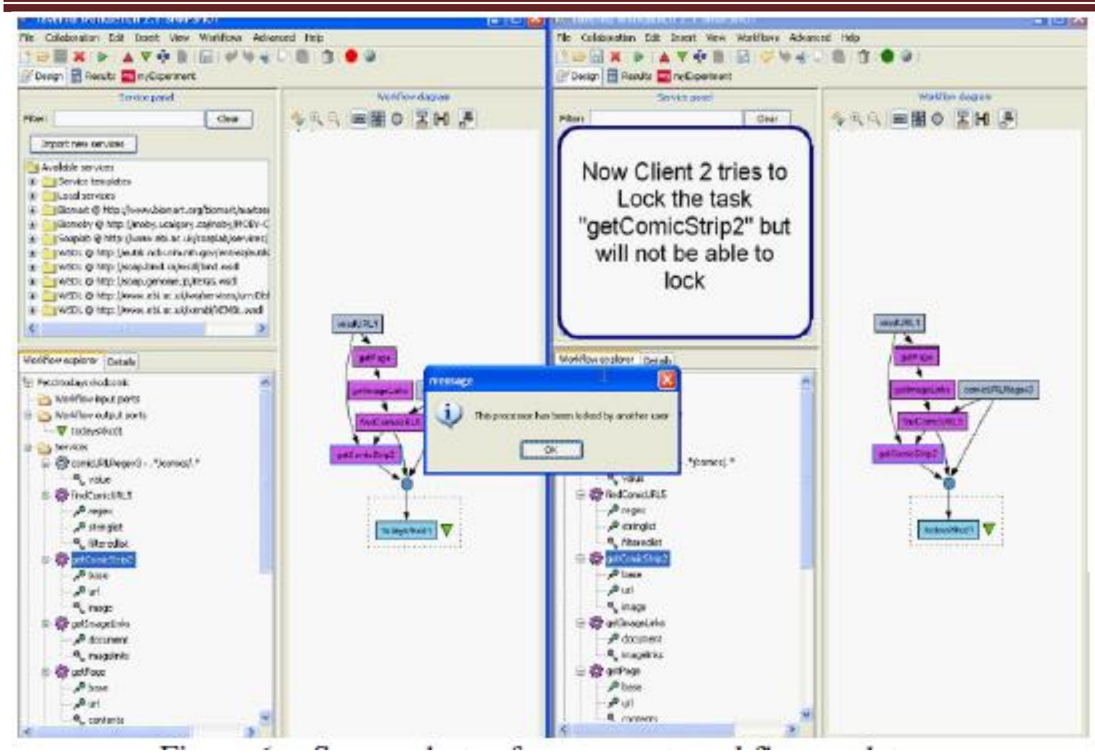
b) Collaboration Transactions

Our locking algorithms facilitate concurrent workflow composition. Actions by each user can be modeled as transactions to ensure concurrency control. We define four basic actions: 1) Insert a data link, 2) Delete a data link, 3) Insert a processor, and 4) Delete a processor. An update action can be modeled as a delete followed by an insert. Thus, all collaborative composition actions can be mapped to database update operations. As a result, we can exploit the concurrency control facility of database management systems to ensure the Serializability of all executions. Bad transactions will be automatically aborted. We are working on an exception handling facility here, which is out of the scope of this paper. After a user update is successfully committed, all collaborators will be notified, so that each collaborator can have the most-up-to date workflow.

2. System Design and Experiments

We built a collaboration pattern template library. The basic building blocks are 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. We built a central server supporting all workflow collaborations. Workflow evolution provenance and collaboration provenance are stored in a shared database on the server. Each collaborator may store an intermediate version of the workflow on the local machine, but all committed activities are stored at the server, in order to support asynchronous collaboration where collaborators may decide to work on the shared workflow at preferable time. We consider four options for selecting database systems: native XML, relational, XML-relational, and RDF. Currently we use a relational database because it is the preferred choice for Taverna.

ISSN: 2320-1363



Now Client 2 tries to Lock the task "getComicStrip2" but will not be able to lock

message

This processor has been locked by another user

Fig.4. Screen Shot of concurrent workflow updates.

It shows a snapshot of our Confucius system supporting concurrent workflow composition. To ease illustration, we show two screens (left and right) representing two scientists running two client versions of Confucius on two distributed machines. Here we use remote desktop feature of Windows to show the two screens together. When a scientist write locks a task on the shared workflow, the other scientist cannot update the task due to our concurrency control.

Concurrency Control Experiments

Our experiments focus on testing the throughput of the Confucius system by varying the number of collaborators. The throughput is defined as the number of successful task updates per minute. The average throughput is calculated for each collaboration group size of N (10, 20, …, 100). For each group size, the experiment is repeatedly performed 10 times with the average calculated. We also monitor the number of failed task updates performed per minute to show the trend of update conflicts as the number of collaborations increases.

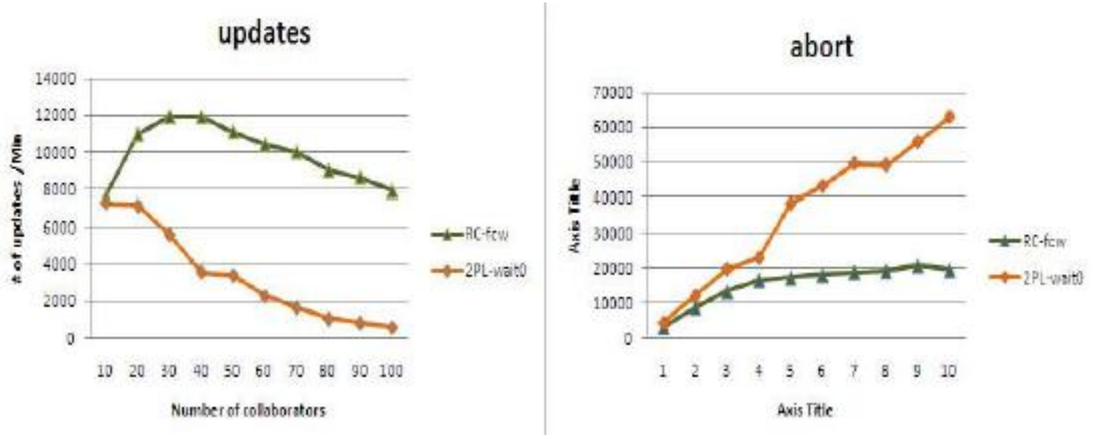


Fig.5. Test result of throughputs and failed task updates.

It shows how the number of successful task updates per minute for varying number of collaborators, from 76,242 task updates per minute for 10 collaborators, to 79,015 task updates per minute for 100 collaborators. We can see that the collaboration productivity (represented by the throughout) is steadily increased as the number of concurrent scientist collaboration increases, reaching a maximum of 119,207 task updates per minute at a group size of 30. Afterwards, the group productivity starts to decline due to the increase of conflicts that leads to abortion.

In summary, from a concurrency control point of view there exits an optimal number for the group size that optimizes the productivity of the system. How to increase such a number, which is the ideal speedup of productivity, is an interesting and challenging open research problem. We plan to further study this problem in our future research.

3. Conclusion

In this article we presented out ongoing work on establishing collaboration protocols to support collaborative scientific workflow composition. Our framework includes a collaboration ontology associated with a set of collaboration patterns, primitives and constructs and a number of concurrent control mechanisms to support concurrent collaborative workflow composition. Based on the ontology, we plan to enhance collaboration provenance management performance by extending our previous work on provenance to support efficient collection, storage, and querying of collaboration provenance, leveraging existing relational, RDF, and XML database techniques. Furthermore, we plan to conduct more experiments to study the effects of tuning various parameters on concurrent productivity.

References

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

[2] LSST, "Large synoptic survey telescope," 2009, Available from: http://www.lsst.org/lsst/science.

[3] LHC, "Large Hadron Collider," 2010, Available from: http://public.web.cern.ch/Public/en/LHC/Computing-en.html.

[4] B. Ludäscher, "Scientific workflows: cyberinfrastructure for e-Science," Proc. of PNC, Oct. 19, 2007, Berkeley, CA, USA, pp.

[5] Y. Gil, E. Deelman, J. Blythe, C. Kesselman, and H. Tangmunarunkit,"Artificial intelligence and grids: workflow planning and beyond", IEEE Intelligent Systems, Jan.-Feb., 2004, 19(1): pp. 26–33.

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

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

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

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

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

[11] S. Wuchty, B. Jones, and B. Uzzi, "The Increasing Dominance of Teams in Production of Knowledge", Science, 2007, 316: pp. 1036- 1039.

[12] N.R. Council, "Facilitating Interdisciplinary Research". 2004, National Academies Press, Washington DC, USA.

[13] G. Bell, T. Hey, and A. Szalay, "Beyond the Data Deluge", Science, 2009, 323(5919): pp. 1297-1298.