Scientific Workflow: Modeling Methods and Management System

Xiaoya Sun¹, Liang Hu² and Xilong Cheª

¹College of Computer Science and Technology, Jilin University, Changchun 130012, China
²College of Computer Science and Technology, Jilin University, Changchun 130012, China
ªCorresponding author: chexilong@jlu.edu.cn

Abstract. With the development of network technology and resource virtualization platform such as e-Science, workflow technology has expanded its coverage from business management to scientific computing. It cooperates with scientists from different fields to achieve complex data processing. The workflow model stands as the core of the workflow management system, and its quality directly affects the execution ability of the whole system. Workflow modeling technology has been widely studied, and many kinds of modeling methods have emerged. This paper proposes a modeling method based on spatio-temporal topology which fuses data flow and control flow together, and presents a workflow management architecture.

1. Introduction
 Workflow technology is originated in the field of office automation in the mid-1970s. The purpose was to automate the business process in the enterprise. The workflow was decomposed and fell into various task roles, which were performed according to pre-defined rules, as well as monitored to improve the management of the enterprise.

With the rapid development of the Internet era, the processing of massive data has become a problem that each industry has to face. In order to adapt to the efficient processing of services in the new mode in various fields, to reduce the large amount of calculation in each link, and to avoid the errors and time consumption of information on manual transmission, workflow technology has been deeply studied and widely used. Today, workflow has not only been applied to traditional business transaction process such as the office automation, document management, database technology and other, but also played an important role in the scientific computing like atmosphere, ocean and geophysics.

Workflow concerns on how to automate a series of activities, whose model description ability directly affects the quality of the workflow system. At present, most workflow models are based on the execution order of activities as the only dependency on workflow modeling. They can’t reflect the generation, transmission and consumption of data between activities.

In response to the limitations of traditional model expression, a workflow model based on spatio-temporal topology is proposed to fuse data flow and control flow together, and the architecture of workflow management system is given in accordance with the model.
2. Modeling method based on UML Activity Diagram

UML is a visual modeling language. It defines an abstract language which is used to describe the structure and behavior of software system. It has a standard graphical notation for building model element view. The activity diagram represents the flow of control from one activity to another, it captures the activities that will be performed based on changes in the state of the object. A next in turn activity will be executed immediately after the completion of current activity in the activity diagram.

UML activity diagram provides a wealth of modeling elements for workflow modeling, as is shown in Fig 1[2]:

![Activity Diagram Elements](image)

- **State**: Represents a state of the atomic task or event.
- **Activity Node**: Expresses the atomic task or event.
- **Start Node**: Represents triggering the entire process.
- **Termination Node**: Represents the ending of the entire process.
- **Or-Connection/Branch**: Represents a conditional branch and selects the chosen branch.
- **And-Connection/Branch**: Represents all branches executed simultaneously.

![Workflow Model](image)

A simple modeling example is shown in Fig 2. The example workflow intends to show the process form of UML. It begins with a company clerk calling the customer to determine a meeting. If the meeting location is within the company, the clerk needs to arrange a meeting room for the meeting; otherwise, the clerk needs to use a notebook to prepare the meeting minutes. The clerk and the
customer meet at the appointed time and place. If an agreement is reached at the meeting, the clerk needs to record it in the memo.

3. Scientific Workflow Management System

3.1 Function of workflow management system

A workflow management system is a computer software system that supports the efficient execution of enterprise business processes and monitors their execution status. A workflow management system can provide the following three functions [3]:

Model establishment function: it mainly considers the definition and modeling function of workflow process and related activities.

Model operation function: it executes the workflow process in certain operating environment, completes the sorting and scheduling functions of the activities in each process.

Human-computer interaction function: it realizes the interaction between user and IT tools during the execution of various activities.

Fig 3 shows the relationships among the three major functions of the workflow management system:

![Fig 3. The main functions of the workflow management system](attachment:image)

3.2 Scientific Workflow Management System

Scientific workflow (SWF) is originated from traditional workflow technology. With the development of resource virtualization platform such as network technology and e-Science, the data analysis process and data analysis environment in scientific research are becoming increasingly complex. The scientific computing process involves the cooperation and collaborative development of scientists from different fields in different organizations. In order to reduce the effort of researchers on calculation-related work, scientific research organization urgently needs a scientific workflow management platform to manage complex scientific computing processes. SWF technology is born in this context and serves as a technology platform for the promotion and guarantee of collaborative development [4]. It provides scientists with a supporting platform for process definition and automatic execution of scientific calculation.

Definition 3.1 Scientific Workflow (SWF) [5]: a scientific workflow is a process that helps scientific experiment through automated or semi-automated methods. Data flow is used to combine workflow tasks to complete data processing and analysis, to achieve the purpose of accelerating scientific calculation.

Definition 3.2 Scientific Workflow Management System (SWFMS) [5]: it is a system that defines, modifies, manages, monitors and executes the SWF. The execution sequence of the tasks is controlled through the formal description of the workflow logic.
4. Workflow Model and System Architecture Proposed

There are mainly two kinds of flows in the workflow management system: data flow and control flow. Control flow describes the execution sequence between activities, it schedules and controls the automation of business activities from the top layer; data flow provides data services for business activities from the bottom layer to ensure the execution of the control process [6]. The interaction between workflow activities is embodied in the interaction of data objects in the activity instance. Data flow occurs with the execution of the control process.

At present, the workflow models adopted by most companies only reflects the control flow in the process. They can’t reflect the generation, transmission and consumption of data between activities. In order to address this issue, this paper proposes a workflow model based on spatio-temporal topology, to achieve a more complete, comprehensive and three-dimensional representation of the workflow.

4.1 Workflow model based on spatio-temporal topology

4.1.1 Concept of spatio-temporal topology

Spatio-temporal topology means that each activity in the workflow is established from both the temporal and spatial levels simultaneously so that the type of relationship is no longer a single control flow. We use five routing structures to define a complete workflow model based on spatio-temporal topology, it has three types of relationships: enclosure relationship, sequence relationship, and data dependency.

UML activity diagram has simple graphic symbols and is easy to expand. For complex processes, it does not have too many model elements and logical relationships as Petri net. Therefore, we make extension on the UML activity diagram based on spatio-temporal topology.

4.1.2 Description of routing structures

Sequential routing: it means that activities are executed in sequence according to certain rules.
Selection-split routing: it means that the activity has more than one conditional branch, and it selects the ones that satisfy the execution condition.
Selection-join routing: an activity has several precursor branches, and one or more of them are activated. The process can be continued when all activated branches have been executed.
Parallel-split routing: it means that all successor branches meet the conditions, regardless of the execution order.
Parallel-join routing: it has multiple parallel branches converging to one node, the successor task can be executed when each branch has finished.

4.1.3 Description of activity relationships

There are three kinds of activity relationships, as is illustrated in what follows:

Enclosure relationship: in the business process, some activities are sub-activities of other activities, such can be defined as enclosure relationships. Assuming that activity A is a sub-activity of activity B, a sufficient condition for activity B to perform is the completion of activity A, but activity A is not necessarily the only condition that triggers activity B. In addition, some complex loop structures can also be encapsulated into a sub-process, represented by enclosure relationships.

Sequence relationship: sequence relationship is modeled based on the start time and end time of each activity. Sequence relationship is divided into time-related relationship and time-unrelated relationship. Time-related one means that the latter activity needs to wait for the completion of the previous activity; time-unrelated one means that the latter activity starts later than the previous one, regardless of whether the previous activity is completed.

Data dependency: data dependency considers the data flow between activities. If the output of one activity is the input of another activity, there is a data dependency between them. Data dependency is divided into strong data dependency and weak data dependency. For activity A and activity B, a strong data dependency means that the execution of activity B must take the output data of activity A as input.
data; a weak data dependency means that whether the output data of activity A is necessary for activity B depends on the attribute value and status of activity B.

4.1.4 Modeling of data flow
Each activity v involves a series of data items, including input data and output data. The output data of an activity is the input data of another. In which the starting activity only has output data and the ending activity only has input data. Strong data dependency can be further divided into two types [7]:

- Execution data dependency: the input data of activity vj is the output data of activity vi, and the activity vj will start to execute only if it obtains the output data of activity vi.
- Selection data dependency: the activity vi is executed when the input data of activity vi satisfies a certain condition, for example, vi is in a certain range.

The activities are modeled in pairs according to the data dependency relationships, as is shown in Fig 4. The data items on the link represent the output data of the former activities and the input data of the latter activities. Solid line indicates the execution data dependency, and dashed line indicates the selection data dependency.

![Fig 4. Data dependency between activities](image)

The overall system design of proposed workflow management system is shown in Fig 5:

![Fig 5 Workflow management system architecture](image)

The entire system contains three modules: user module, workflow engine, and application module.

1. User module: the user module visualizes the workflow instance to the customer. The user interface can create the workflow model and show the executive results of scientific workflow. The administrator manages the right of users. The workflow server provides an interface service that links user and the workflow engine.

2. Workflow engine: the workflow engine is the core of the workflow system. It interprets process definition files, extracts relationships between entities, visualizes workflow model to user, initializes workflow instances, and evaluates the workflow performance [8]. The workflow engine consists of seven parts, which are described as follows:
Data loading: it loads process definition data from the database into workflow engine, and transfers to the XML generator.

XML generator: the data entered by the user may be in many formats, it converts data to XML format according to the corresponding mapping rules. If the data itself is XML, skip this step.

XML parser: its function is to parse the xml file. It extracts the nodes and the relationships between them.

Process instantiation: the nodes and relationships extracted from each XML file are organized according to certain rules, to form a complete workflow model.

Workflow visualization: it visualizes the workflow model and presents to the user.

Workflow evaluation: it evaluates the generated workflow model to determine if the model meets the executable standard.

Workflow storage: it stores the generated workflow model, which can be called directly during next-time execution.

3. Application module: workflow management system interacts with external resources through two ways: client application and direct invocation application. In short, the former one requires user participation, and the latter one does not.

There are three data components feeding the data for the operation of various modules in the workflow management system, namely User data, process definition data, and model control data. User data component stores users' personal and access control information. Process definition data component stores workflow data that needs to be processed in scientific computing. Model control data component stores related data defined in the workflow model, such as routing rules, representations of relationships between nodes, etc.

5. Conclusion
This paper proposes a scientific workflow model based on spatio-temporal topology, which fuses data flow and control flow together and bridges the gap of the traditional model in expressing the activity relationship. By adding sequence relationship, enclosure relationship and data dependency, the model expression is more complete and is extended in a three-dimensional space. This paper also presents the overall architecture of the workflow management system on accordance with the model. The scientific workflow system lacks the capability of multi-level abstract expression, which can't meet the diverse needs of current application/users simultaneously. Our future research will focus on these deficiencies.

Acknowledgement
This work is funded by: National Key R&D Plan of China under Grant No. 2017YFA0604500, National Sci-Tech Support Plan of China under Grant No. 2014BAH02F00, and by National Natural Science Foundation of China under Grant No. 61701190, and by Youth Science Foundation of Jilin Province of China under Grant No. 20180201103GX.

References
[1] Guangzhi Chen, Yi Pan, Lei Li. A Survey of Workflow Modeling Technology and Its Research Trends[J]. Computer Science, China,2014, 41(s1):11-17.
[2] Modelling F W, Brinksma P H, Eshuis H. Semantics and Verification of UML Activity Diagrams for Workflow Modelling[J], 2017.
[3] Workflow T, Coalition M. Workflow Management Coalition Interface 1 : Process Definition Interchange Process Model[J]. 1999.
[4] Gideon Juve, Ann Chervenak, Ewa Deelman, et al. Characterizing and profiling scientific workflows[J]. Future Generation Computer Systems, 2013, 29(3):682-692.
[5] Ailamaki A, Ioannidis Y E, Livny M. Scientific workflow management by database management[C]// Scientific and Statistical Database Management, 1998. Proceedings. Tenth International Conference on. IEEE, 1998:190-199.

[6] Peter P. S. Chen. The entity-relationship model: toward a unified view of data[C]// International Conference on Very Large Data Bases. ACM, 1975:173-173.

[7] Sun S X, Zhao J L. Formal workflow design analytics using data flow modeling[J]. Decision Support Systems, 2013, 55(1):270-283.

[8] Samandi V, Mukhopadhyay D. Proposing an Architecture for Scientific Workflow Management System in Cloud[C]// Springer-International Research Symposium on Computing and Network Sustainability. 2017.