On Scientific Workflow

Dr. Jinjun Chen

Swinburne University of Technology, Australia

E-mail: jchen@ict.swin.edu.au

Web: http://www.ict.swin.edu.au/personal/jchen

Prof. dr. ir. W.M.P. van der Aalst

Eindhoven University of Technology, The Netherlands

E-mail: W.M.P.v.d.Aalst@tm.tue.nl

Web: http://is.tm.tue.nl/staff/wvdaalst/

Scientific workflow is another type of workflow that usually underlies in many complex e-science applications such as climate modeling and disaster recovery simulation. From the perspective of such e-science applications, based on Dieter Cybok's paper in GGF10 there are at least three motivations for scientific workflow: 1) some complex e-science applications often require the creation of a collaborative workflow; 2) many e-scientists lack the necessary low-level expertise to utilize the current generation of underlying computing infrastructure such as Grid toolkits; 3) workflow specifications can be reused, modified and shared once they are defined. With these motivations, specific requirements for developing a scientific workflow management system need to be identified. The requirements may cover several aspects such as data or computation intensity and lifecycle management of participating services. The investigation of such requirements may be carried out from the following two points of view.

1) Those requirements for scientific workflow are also typical in business workflow. Scientific workflow is just another type of workflow. By nature, it should have some features in common with business workflow. Business workflow has been under investigation for more than two decades. There are many techniques which have been developed and many scholars are working in the area. By identifying those common requirements such as control flow modelling, even-driven analysis and large-scale collaboration, we can try to adapt corresponding techniques from business workflow to scientific workflow rather than develop them again. Some of those requirements such as interactive

steering may not be well supported by business workflow for the time being. However, since they are also needed by business workflow, business workflow scholars may probably be working or will work on them because business workflow has been an area for a long time. In such situation, we may go along scientific workflow domain in parallel with business workflow domain. But more importantly, we should notice the recent advances in business workflow domain to see whether new techniques have been developed which can be adapted to scientific workflow.

2) Those requirements for scientific workflow cannot be seen or are not typical in business workflow. This point should be more important as it makes the necessity of the name of “scientific workflow”. If all requirements for scientific workflow are also typical in business workflow, then even if some of them cannot be well supported by business workflow techniques for the time being, there may not be too much further research for us to do as we can simply apply existing or upcoming techniques of business workflow to scientific workflow. Therefore, we need to identify scientific workflow specific requirements such as computation or data intensity and dynamic resource allocation, scheduling and mapping to underlying distributed infrastructure such as grid computing environments. For example, a scientific workflow normally contains a large number of data or computation intensive activities. Accordingly, a scientific workflow management system needs to accommodate a large amount of computation and transfer a huge amount of data between participants (grid services if supported by a grid environment). Decentralised data transfer might be a good way such as in a peer-to-peer fashion, i.e., directly between participants rather than via the scientific workflow engine. Corresponding techniques for modelling interfaces of supporting services will also be needed.

Many efforts have been made on scientific workflow from scientific domain. For example, GGF10 and its special issue in Concurrency and Computation: Practice and Experience were early efforts. In addition, a special issue in International Journal of High Performance Computing Applications was another effort. A special group in GGF (now OGF) was set up and is an ongoing effort. Scientific workflow is also a focus of IEEE TCSC Technical Area on Workflow Management in Scalable

Computing Environments which is established recently (http://www.ict.swin.edu.au/personal/jchen/tcsc/WMSCE.htm). Some relevant conferences and workshops have been held or are being run. For example, WaGe2007 (2nd International Workshop on Workflow Management and Applications in Grid Environments - http://www.ict.swin.edu.au/personal/jchen/WAGE/WAGE07.htm) will be running during August 16-18, 2007, in Urumchi, Xinjiang, China. Another workshop called GPWW2007 (3rd International Workshop on Grid and Peer-to-Peer based Workflows - http://www.ict.swin.edu.au/conferences/gpww/2007/) will be held on Sept. 24, 2007 in Brisbane, Australia. This workshop is in conjunction with the 5th International Conference on Business Process Management (BPM 2007). There are also some other workshops such as WSES07, SWF2007 and WORKS07, and some related projects such as SwinDeW-G, Gridbus Workflow and Pegasus. SwinDeW-G is a decentralised grid workflow management system (http://www.ict.swin.edu.au/personal/jchen/SwinDeW-G/System_Architecture.pdf) in which the workflow execution and information interaction between participants are performed in a P2P fashion. SwinDeW-G is being ported into a grid infrastructure called SwinGrid. The web links of some related conferences, workshops and projects can be found at the Technical Area website.

With the efforts from scientific domain, gradually business workflow domain is also paying more and more attention to scientific/grid workflow. For example, the group of Prof. W.M.P. van der Aalst in Eindhoven University of Technology (TU/e) in The Netherlands has achieved a lot of experience in process modelling, analysis and enactment. The workflow patterns (www.workflowpatterns.com) have become a standard way to evaluate languages and the workflow management system. YAWL is one of the most expressive and mature open-source workflow systems available today (www.yawl-system.com). Moreover, they have been specialising in process analysis. Using Petri nets as a theoretical foundation, they have been able to analyse a variety of real-life process models ranging from BPEL and workflow specifications to the entire SAP reference model. Moreover, in recent years, they have focused on the analysis of processes based on system logs. The ProM framework developed at TU/e provides a versatile toolset for process mining (www.processmining.org), which seems to be particularly useful in a grid environment. In overall terms, they are trying to bring such knowledge together with grid computing in order to make further progress in both areas. Specifically, they are trying to apply their ample knowledge to modelling grid applications, analysing grid workflow models and grid system logs, and building a process-aware grid infrastructure. They use a mixture of Petri nets and UML modelling to build formal/conceptual models for grid computing. Also, using Petri-net-based techniques, they analyse different mechanisms used in grid workflows in order to transfer correctness notions such as soundness to grid workflows. Since in a grid environment many events are logged and the performance of the system is of the utmost importance, they are interested in applying their process mining techniques to the domain. By linking a fundamental enabling technology for the grids (Globus) to a powerful process engine (YAWL) and state-of-the-art analysis tools (ProM), they obtain an interesting environment for experimentation towards building a process-aware grid infrastructure.

Combining efforts from both scientific domain and business workflow domain might be able to provide a balanced way to exploring scientific workflow. This is one of the motivations for the IEEE TCSC Technical Area on Workflow Management in Scalable Computing Environments (http://www.ict.swin.edu.au/personal/jchen/tcsc/WMSCE.htm). Since the technical area is located in IEEE TCSC, it is automatically associated with scientific domain. To grab more attention and efforts from business workflow domain, its steering committee consists of several world-class scholars from the business workflow/process area.