Future Generation Computer Systems 25 (2009) 525–527

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Editorial

Special section on workflow systems and applications in e-Science

Zhiming Zhao [a,*], Adam Belloum [a], Marian Bubak [a,b]

[a] University of Amsterdam, Kruislaan 403, 1098SJ, Amsterdam, The Netherlands
[b] Institute of Computer Science, AGH, Mickiewicza 30, 30-059 Krakow, Poland

ARTICLE INFO

Article history:
Received 21 August 2008
Accepted 29 October 2008
Available online 12 November 2008

Rapid development of Internet and Grid technologies has greatly extended the number of resources which scientists can apply for research. More and more scientific advances are achieved via experiments which consist of large volumes of computing tasks and data. Scientific Workflow Management Systems (SWMS) are a key technology to integrate computing and data analysis components, and to control the execution and logical sequences among them. By hiding the complexity in an underlying infrastructure, SWMSs allow scientists to design complex scientific experiments, access geographically distributed data files, and execute the experiments using computing resources at multiple organizations. In this way, domain scientists can effectively use available resources while focusing on the logic of the experiments instead of low level technical details.

During the past few years, scientific workflow systems and applications have attracted enormous research interest. Fig. 1 shows an increasing number of workflow related publications in the journal of Future Generation Computer Systems (FGCS). A number of research foci can be found in recent publications. Running legacy applications on the Grid is essential for the domain scientists to extend execution scenarios of their experiments. Wrapping components in a scientific workflow using a standardized architecture, such as services, improves the reusability of the components and also makes the integration easy. McGough et al. [1] present a service architecture named GridSAM to manage the job submission onto Grid and to manage different abstractions of tasks at runtime. Glatard et al. [2] demonstrate how services can be dynamically grouped and optimized in the workflow enactment and scheduling at runtime. Interoperability between different workflows is essential to realizing cooperation between different scientific experiments. Describing distributed Grid resources using semantic web technologies promotes high-level sharing of components and workflows [3]. The scientific workflow has proven to be a key technology to enhance many research disciplines to manage the experiment processes and automate computation; one of the examples is the data mining application [4].

Several workshops and special issues have been organized to extensively discuss research topics covering the lifecycle of scientific experiments, i.e., the experiment design, execution, data process and publication. These fora not only cover a wide spectrum of interests in the SWMS development and application, as shown in Fig. 1 but also have different foci, e.g., WSES[1] addresses the main challenges in system development, SWF[2] highlights the service oriented architecture in scientific workflows, SWBES[3] foci on using industrial workflow standards in e-Science, and WORKS[4] has clear interests in workflows in data intensive applications. As a follow up of these workshops, several special issues on scientific workflows have been published in the journals of *Scientific Programming* [5], and *Concurrency and Computation: Practice and Experience* [6]. The book *workflows for e-Science* [7] is a good collection of publications about this subject.

Moreover, scientific workflows have also been the main objective of several projects: the Dutch Virtual Laboratory for e-Science (VL-e),[5] the EU funded Knowledge Workflow Grid (K-WfGrid)[6] and the ViroLab projects.[7] The Dutch VL-e project aims at sharing and transferring knowledge in the form of mature experiments encoded using workflow descriptions from

* Corresponding author. Tel.: +31 205257599.
E-mail address: z.zhao@uva.nl (Z. Zhao).

0167-739X/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2008.10.011

[1] International Workshop on Workflow Systems in e-Science, http://staff.science.uva.nl/~zhiming/workshop/wses/.
[2] International Workshop on Scientific WorkFlows, http://www.cs.wayne.edu/~shiyong/swf/.
[3] International Workshop on Scientific Workflow and Business workflow standards in e-Science, http://staff.science.uva.nl/~zhiming/workshop/swbes/.
[4] The workshop on Workflows in Support of Large-Scale Science, http://www.isi.edu/works07/.
[5] http://www.vl-e.nl.
[6] http://www.kwfgrid.net/.
[7] http://virolab.org.

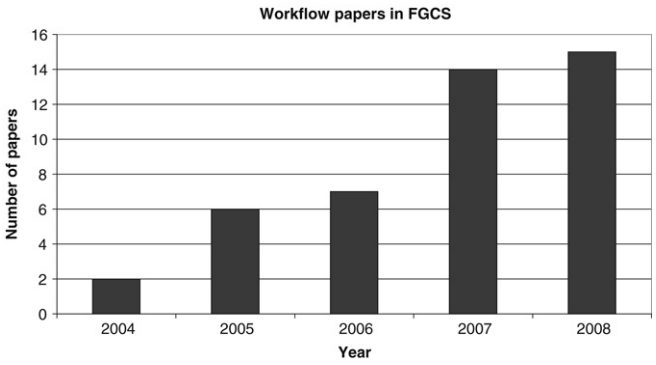

Fig. 1. Scientific workflow related papers in FGCS.

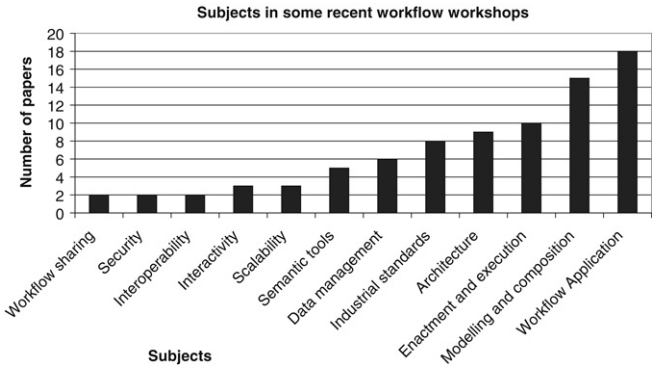

Fig. 2. Subjects in recent workflow workshops (WSES 06,07 and 08, SWBES 06 and 07, SWF 07 and WACS 08).

six different application domains, where different workflow systems are recommended to different application scientists. By abstracting generic components from a specific application, a generic e-Science framework is being developed in the VL-e project. The key research question in the K-WfGrid project was to bridge a gap between the high level workflow description and available components and services. This was done using semantic technology to enable the (semi)automatic composition of the scientific workflow. The main result of the ViroLab project[8] is a prototype of a decision support system for HIV drug resistance, where a script-based language is used to describe experiments, while semantics are applied to compose applications. The ViroLab virtual laboratory is equipped with an advanced provenance subsystem.

This special section is dedicated to workflow systems in e-Science and includes specially selected papers from three workshops: WSES07, SWF07 and SWBES07. The papers address various aspects of scientific workflows including workflow scheduling, provenance, data management, and applications of workflows in e-Science. As shown in Fig. 2, e-Science applications are an important driving force for scientific workflow systems.

Different application requirements and the diversity of experimental methods in different research disciplines result in a collection of workflow tools and implementations which demonstrate a wide variety of capabilities. Deelman et al. [8] review the features and capabilities of existing workflow systems and discuss current research issues in scientific workflows from the perspective of the workflow lifecycle: composition, mapping, execution and publication.

Modeling and composing a workflow is a basic functionality provided by workflow systems. McPhllips et al. [9] argue, from the usability point of view, that a workflow system should have several important properties, such as declarativity, well-formedness, predictability and recordability and they propose an assembly line based on a workflow representation metaphor called COMAD. The workflow system adds variations and realizes languages to describe the control and data flows in applications in a specific set of domains. Many workflow systems provide a text or graphical interface for computing workflows. However, manual composition is not always practical and it is difficult to graphically render large scale and complex workflows. Using semantic technologies for (semi)automatic generation of workflows is an important trend. Mapping an abstract workflow description onto concrete computing and data storage resources, and orchestrating the runtime behavior of workflow processes comprise the basic execution procedures of a scientific workflow.

Decoupling the model of computation from the implementation of the workflow engine allows composing workflows that include multiple models of computations; however, it also requires user knowledge in understanding different combinations of the model of computation. Goderis et al. [10] discuss the characteristics of different workflow execution models and analyze compatibility between computational semantics of different execution models provided by the Ptolemy system. It is an important step towards providing high-level workflow composition support. Most of the SWMS are historically driven by applications in specific domains, e.g., bioinformatics, high energy physics and astronomical observations.

E-Science applications require the investigation of the common characteristics of domain-specific systems and their implementation as part of a generic framework. A number of research projects such MyExperiment[9] and VL-e aim at developing a Grid-enabled generic framework where scientists from different domains can share their knowledge and resources to perform domain specific research. Sharing knowledge and resources requires more interoperability among the major workflow management systems. As more sophisticated solutions are needed to achieve a seamless integration of workflows, approaches such as the workflow bus [11] developed in the context of VL-e present a potential solution to the interoperability problem. Different requirements for supporting domain specific applications are an important driving force for the development of workflow systems. De Roure et al. [12] present a social web based cooperative virtual research environment, called myExperiment, for sharing workflows and experiments. A user can publish workflow descriptions, search for interesting workflows from the other community members, and execute a workflow online. In addition to the Taverna engine, the myExperiment environment also attempts to integrate several other workflow engines. Executing a workflow using multiple workflow engines is becoming steadily more feasible since the emergence of Service Oriented Architecture (SOA) as a potential architectural model to implement new generations of SWMS. A number of workflow research projects worldwide have started to move towards SOA, the monolithic implementation of SWMS is being replaced by two separate entities: the workflow composer used at the design time and the workflow engine responsible for the run time execution. Implementing the workflow engine as a standalone component and using standard technology to invoke it, such as Web services, is a potential approach to distribute a workflow among multiple workflow engines [13].

Recording the interactions and data evolution in a complex experiment, which is called data provenance, is crucial to trace the history of the experiment processes and to reproduce the

[8] http://www.virolab.org/.

[9] http://www.myexperiment.org.

experiment results. The provenance service is becoming an important part of several workflow systems and attracting much interest in the research community. Data storage and semantic technologies based annotation and queries are playing a key role. Wang et al. [14] explore how to provide provenance to the transactions of workflow activities. Shao et al. [15] discuss the discovery of critical workflows from provenance data.

This special section also includes a workflow application in Earth Systems simulation using Microsoft Workflow Foundation (WF) technology. The WF technology provides a rich set of features to support the authoring and execution of workflows, tracking services that enable the monitoring of a running workflow, and state persistence services that allow workflows to be recovered and resumed upon failure. Fairman et al. [16] discuss how Microsoft Visual Studio IDE and the Windows Presentation Foundation which delivers a browser based client interface are used to design and manage a GENIE workflow.

In development of a workflow management system, one faces two important challenges. On one hand, domain specific experiments require customized solutions in workflows for particular problems; on the other hand, to enable knowledge transfer and information sharing between different domains, a generic workflow solution is also required. A successful workflow system should not only have a mature conceptual design and engineering but, more importantly, it should effectively enhance real life applications. The usability of a workflow system is essential to make it useful in the day-to-day activity of application domain scientists: not only suitable interface for composing and executing workflow, but also a set of user oriented tools for viewing, moving and processing data and for the provenance of the workflow and reproducing the results. Sharing the knowledge in meaningful workflows is becoming an important requirement for e-Science frameworks. It is necessary to integrate different technologies, e.g., semantic annotation and searching, and collaborative working facilities, with a scientific workflow system. We can enumerate a number of foci for the future development of workflow systems: workflow sharing and discovery, provenance, human in the loop workflow execution, and workflow interoperability. These extend the list of requirements formulated recently in [17].

Acknowledgments

We would like to express great appreciation to all the program committees, referees and colleagues of the workshops and those who have supported the special section. We also specially acknowledge support from the Dutch Virtual Laboratory for e-Science project.

References

[1] A. Shikta Das, Stephen McGough, William Lee, A standards based approach to enabling legacy applications on the grid, Future Generation Computer Systems 23 (2008) 731–743.
[2] Tristan Glatard, Johan Montagnat, David Emsellem, Diane Lingrand, A service-oriented architecture enabling dynamic service grouping for optimizing distributed workflow execution, Future Generation Computer Systems 24 (2008) 720–730.
[3] Hong-Linh Truong, Schahram Dustdar, Thomas Fahringer, Performance metrics and ontologies for grid workflows, Future Generation Computer Systems 23 (2007) 760–772.
[4] A. Ping Luo, L. Kevin, Zhongzhi Shi, Qing He, Distributed data mining in grid computing environments, Future Generation Computer Systems 23 (2008) 84–91.
[5] A. Belloum, E. Deelman, Z. Zhao, Scientific workflows, guest editorial, Scientific Programming 14 (2006) 171.
[6] Geoffrey Charles Fox, Dennis Gannon, Special issue: Workflow in grid systems, Concurrency and Computation: Practice and Experience 18 (10) (2006) 1009–1019.
[7] I.J. Taylor, E. Deelman, D.B. Gannon, M Shields (Eds.), Workflows for e-Science, Springer, 2007.
[8] Ewa Deelman, Dennis Gannon, Matthew Shields, Ian Taylor, Workflows and e-science: An overview of workflow system features and capabilities, Future Generation Computer Systems 25 (5) (2009) 528–540.
[9] Timothy McPhillips, Shawn Bowers, Daniel Zinn, Bertram Ludaescher, Scientific workflow design for mere mortals, Future Generation Computer Systems 25 (5) (2009) 541–551.
[10] Antoon Goderis, Christopher Brooks, Ilkay Altintas, Edward A. Lee, Carole Goble, Heterogeneous composition of models of computation, Future Generation Computer Systems 25 (5) (2009) 552–560.
[11] Zhiming Zhao, Suresh Booms, Adam Belloum, Cees de Laat, Bob Hertzberger, Vle-wfbus: A scientific workflow bus for multi e-science domains, in: Proceedings of the 2nd IEEE International conference on e-Science and Grid computing, IEEE Computer Society Press, Amsterdam, The Netherlands, 2006, pp. 11–19.
[12] David De Roure, Carole Goble, Robert Stevens, The design and realization of myexperiment virtual research environment for social sharing of workflows, Future Generation Computer Systems 25 (5) (2009) 561–567.
[13] V. Korkhov, D. Vasunin, A. Wibisono, A. Belloum, M. Inda, T. Breit, M. Roos, L.O. Hetrzberger, Vlam-g: Interactive dataflow driven engine for grid-enabled resources, Scientific Programming 15 (3) (2007) 173–188.
[14] Liqiang Wang, Shiyong Lu, Xubo Fei, Artem Chebotko, H. Victoria Bryant, Jeffrey L. Ra, Atomicity and provenance support for pipelined scientific workflows, Future Generation Computer Systems 25 (5) (2009) 568–576.
[15] Qihong Shao, Peng Sun, Yi Chen, Efficiently discovering critical workflows in scientific explorations, Future Generation Computer Systems 25 (5) (2009) 577–585.
[16] Matthew J. Fairman, Andrew R. Price, Gang Xue, Marc Molinari, Denis A. Nicole, Timothy M. Lenton, Robert Marsh, Kenji Takeda, Simon J. Cox, Earth system modelling with windows workflow foundation, Future Generation Computer Systems 25 (5) (2009) 586–597.
[17] Yolanda Gil, Ewa Deelman, Mark Ellisman, Thomas Fahringer, Geoffrey Fox, Dennis Gannon, Carole Goble, Miron Livny, Luc Moreau, Jim Myers, Examining the challenges of scientific workflows, Computer 12 (2007) 24–32.

Zhiming Zhao studied Computer Science in Nanjing Normal University (NJNU) (1989–1993) and East China Normal University (ECNU) (1993–1996) in China, and obtained a Bachelor and Master of Science degree in 1993 and 1996 respectively. He received a Ph.D. in computer science from University of Amsterdam (UvA) in 2004. He is currently a researcher at UvA. His research interest includes scientific workflow management systems, distributed interactive simulation, e-Science and Grid computing. He has published more than 30 papers in peer reviewed conferences and journals, and chaired several workshops on workflow systems and applications in the context of Int. conf. on ICCS, IEEE e-Science and IEEE CCGrid since 2006.

Adam Belloum is an Assistant Professor at the computer science department of the University of Amsterdam. He received the M.Sc. and Ph.D. degrees from the Compiegne University of Technology, France. He started his research activities in 1992, he first worked in the area of designing parallel computers (VLSI design), where he participated in the design and the implementation of a prototype of a parallel computer dedicated to image processing based on DSP processors. In 1997, he moved to The Netherlands where he was leading the research activities in the area of Web caching until the year 1999. During the last six years, he was heading a small research group at the IvI, UvA, which has developed the first prototype of the run time system of the virtual laboratory toolkit. He published more than 30 articles in international conferences and journals and helped in managing three research projects: JERA project (1997–1999), the Virtual laboratory (1999–2002) and the VLe (2004–2009).

Marian Bubak has M.Sc. degree in Technical Physics and Ph.D. in Computer Science. He is an adjunct at the Institute of Computer Science and ACC CYFRONET—AGH University of Science and Technology, Krakow, Poland, and the Professor of Distributed System Enginnering at the Universiteit van Amsteram. His research interests include parallel and distributed computing, grid systems, and e-science; he is author of about 200 papers in this area, co-editor of 15 proceedings of international conferences and the Associate Editor of FGCS Grid Computing. In the CrossGrid Project he was the leader of the Architecture Team, the Scientific Coordinator of the K-WfGrid, the member of the Integration Monitoring Committee of the CoreGRID, and the WP leader in ViroLab and GREDIA EU IST projects.