Guest Editors' Introduction

Incorporating Scientific Workflows in Computing Research Processes

Shantenu Jha
Rutgers University, Brookhaven National Laboratory

Scott Lathrop
Shodor Education Foundation, Inc., National Center for Supercomputing Applications, University of Illinois

Jarek Nabrzyski
Center for Research Computing, University of Notre Dame, Lawrence Berkeley National Laboratory

Lavanya Ramakrishnan
Lawrence Berkeley National Laboratory

We are pleased to present this special issue on "Incorporating Scientific Workflows in Computing Research Processes" with a goal of realizing a significant increase in the development and usage of high-quality workflow tools, environments, and methods in the scientific community.

This special issue aspires to increase awareness of the benefits of workflows to enhance computational and data-enabled research and to foster the exchange of lessons learned and good practices that can benefit the community. The issue highlights some of the activities and approaches that are underway in the scientific workflow community.

Scientific workflows evolved as a way to manage computation on High Performance Computing (HPC) and distributed systems. Early workflow efforts started as domain-specific efforts that managed directed acyclic graphs on high-performance and distributed systems. They considered the systems and the applications as black boxes and focused on distributed resource management, workload, and execution management. In the mid-2000s, there were efforts to provide taxonomiesm,[1] classifications[2] for workflow systems, and a perspective of the field.[3]

The papers included in this special issue constitute a subset of the many workflow challenges and systems that exist in the community, but nonetheless represent the breadth of the systems and capabilities that are available. The special issue also spans a swath of time—from tools that have been around for multiple decades as well as more recent innovations. The six papers selected for the special issue cover a range of topics in workflows and workflow systems, and can be classified into three broad areas: 1) a description of specific workflow tools; 2) experience and investigation into social and technical aspects of

Digital Object Identifier 10.1109/MCSE.2019.2917987
Date of current version 24 June 2019.

Published by the IEEE Computer Society

collaborative computational science; 3) development methodology and tools in response to representative and specific technical problems.

The papers by Oliver and Deelman are about specific tools—Cylc and Pegasus, the former being a tool primarily used in climate, weather, and weather-driven forecasting systems; the latter being a workflow tool used by several domains. In addition to providing technical details of the tools and how they are used, the articles recount the experiences and challenges in evolving tools over the long haul.

The papers by Altintas and Catlin represent different approaches taken to managing the collaboration and communication of a cross-disciplinary team, and the role of tools in the process. For example, Altintas *et al*. investigate interaction methodologies and software services for effective workflow-driven collaborations. Catlin *et al.* share their preliminary sociotechnical experience in developing the DEEDs platform.

The papers by Turilli and Yildiz highlight viewpoints and solutions to specific questions that are increasingly important for the workflow community. Turilli *et al*. investigate the design and development of middleware building blocks as scalable and sustainable approaches to support diversity of requirements as well as software complexity. Yildiz *et al*. outline some of the general challenges in developing workflows with distinct levels of hierarchy, programming, and execution models in the context of integrating task-based and *in-situ* workflows.

OPEN CHALLENGES

The six papers touch upon an impressive range of problems that are drawn from the science of workflows and the engineering of workflow tools. However, as might be expected by a single special issue and a collection of papers that were asked to emphasize experience and practical tools, over formal considerations, several important technical aspects, and underlying research problems that shape the landscape of workflows and workflow systems are not covered.

In the last decade, the growth of experimental and observational data has ushered an era of data-intensive computing that is resulting in additional innovations in scientific workflows. The need to support data analytics workflows, streaming workflows, and human-in-the-loop workflows has been recognized. More recently, scientific workflows have been recognized as a cornerstone that is needed to enable next-generation scientific discoveries and is beginning to be considered as critical in the design of end-to-end systems.[4] This special issue does not touch on these newer workflows and related topics including reproducibility, provenance, automation, and data-intensive frameworks. We anticipate that these topics will be the focus of future discussions and papers.

Furthermore, important functional areas like workflow monitoring, fault-tolerance, and performance portability of workflows are not specifically discussed in this special issue, but are active areas of advances in workflows and system research. Finally, even though several of the papers are motivated by HPC workflows, many aspects central to HPC workflows—e.g., scheduling and resource allocation and management—as well as a detailed discussion of heterogeneous architectures, would benefit from specialized attention via ongoing discussions and the publication of additional papers.

Currently, it is difficult to select a workflow tool that meets the requirements of a project. Large projects can test a subset of workflow systems to find the ones that seem to work best for them but they cannot exhaustively test all systems; small groups generally pick a system based on perceived simplicity, or other partial but possibly suboptimal considerations. There is a need for a structured approach to the selection of suitable, if not optimal, workflow approaches and tools. This was highlighted as one of the most significant challenges associated with workflows at a recent Blue Waters Workflows workshop.[7]

CONNECTING WITH THE COMMUNITY

It is important that information exchange and collaboration between domain and computer scientists is cultivated. There is a need for fora and platforms in which domain and computer scientists may exchange their ideas for the evolution of workflow environment capabilities that can address current and evolving scientific requirements and needs.

Advances in workflows, workflow tools, and associated issues are presented and discussed at diverse workshops and conferences, including the ISC'xx,[5] SC'xx,[6] and IEEE e-Science conference series. The Workflows in Support of Large-Scale Science (WORKS) workshop held in conjunction with the SC'xx conference series is a noteworthy forum for these exchanges.[8]

Even though workshops provide a forum for information exchange, we believe that fora that are more interactive and crafted toward particular sciences are needed. Bootcamps and hackathons provide interactive sessions that enable researchers to understand systems capabilities and confidently select the appropriate system(s). These interactive and engaging formats also enable better opportunities for capturing requirements and can help yield the development of better tools, systems, and environments.

There are multiple options to learn more about workflows. An extensive, although incomplete, list of workflow systems is available,[9] many of which are available at little or no cost to research teams, and a number of them utilize an open source software approach. Overviews of a number of specific workflow systems have been presented through the series of Blue Waters Webinars,[10] for which the recordings of the presentations are publicly available. We welcome other workflow developers and science teams to share information via this forum.

SUMMARY

We hope that this special issue will provide a timely contribution to information exchange and collaboration among domain and computer scientists. We believe that the special issue papers will benefit the scientific workflow community, including both users and developers, by allowing them to build on the experiences and lessons learned. We hope there will be greater sharing and synergy between different workflow efforts, and thereby contribute to increased usability for the research community. Finally, we hope this special issue will induce many more computational and data-intensive research teams to benefit from the utilization of innovative workflow resources to reduce the time to scientific discoveries.

REFERENCES

1. J. Yu and R. Buyya, "A taxonomy of workflow management systems for grid computing," Grid Comput. Distrib. Syst. Lab., Univ. Melbourne, Parkville, VIC, Australia, Tech. Rep. GRIDS-TR-2005-1, Mar. 2005.
2. Y. Han, A. Sheth, and C. Bussler, "A taxonomy of adaptive workflow management," in *Proc. Conf. Comput.–Supported Cooperative Work*, 1998.
3. I. J. Taylor, E. Deelman, D. B. Gannon, and M. Shields, *Workflows for E-Science: Scientific Workflows for Grids*. New York, NY, USA: Springer, 2014.
4. E. Deelman *et al.*, "The future of scientific workflows," *Int. J. High Perform. Comput. Appl.*, vol. 32, no. 1, pp. 159–175, Jan. 2018. [Online]. Available: https://doi.org/10.1177/1094342017704893
5. [Online]. Available: https://www.isc-hpc.com/
6. [Online]. Available: http://supercomputing.org/
7. Blue Waters Workflows Workshop, Aug. 9–10, 2016. [Online]. Available: https://sites.google.com/a/illinois.edu/workflows-workshop/home
8. [Online]. Available: http://works.cs.cardiff.ac.uk/
9. Workflow systems. [Online]. Available: https://github.com/common-workflow-language/common-workflow-language/wiki/Existing-Workflow-systems
10. Blue Waters webinar series. [Online]. Available: https://bluewaters.ncsa.illinois.edu/webinars