Report on SciFlow 2006: The IEEE International Workshop on Workflow and Data Flow for Scientific Applications

Brian F. Cooper
College of Computing
Georgia Institute of Technology
cooperb@cc.gatech.edu

Roger Barga
Microsoft Research
barga@microsoft.com

1. Introduction

Computation has been described as the "third leg" of science, along with theory and experimentation. Certainly, modern information systems are vital to managing and processing the huge amounts of data produced by simulations and experiments. However, existing tools are only now beginning to catch up with the needs of today's scientists, and much more needs to be done to support the computational needs of tomorrow's scientists. In particular, scientists still need effective tools to deal with massive data sets that may be geographically scattered, to apply multiple complex and interacting transformations to the data, and to ensure the quality and repeatability of their computations. The IEEE SciFlow workshop brought together computing researchers who are exploring how to build the next generation of information systems to address these needs. The workshop was held on April 8, 2006, in conjunction with the IEEE International Conference on Data Engineering in Atlanta, Georgia, USA.

The papers presented in this workshop demonstrate the ability of computer scientists and natural scientists to work together to create computer systems that support scientific exploration. The workshop itself was very interactive, with the audience raising many questions for the speakers and different speakers adapting their talks to address points brought up in discussions. This interaction was greatly helped by generous sponsorship from Microsoft, which provided a lunch and cocktail reception so that participants could continue their discussion.

2. Workshop themes

Several overall themes emerged from the discussions. One theme is that although many groups are building systems today, there are still many open research problems and a lack of standard tools for use by scientists. There are several available tools for workflows (such as the Kepler workflow system, Windows Workflow Foundation and components of IBM's WebSphere), but work still needs to be done to adapt many tools to the scientific domain, and make them usable by non-computer scientists. In addition to tool development, important research problems include:

- Applying "general purpose" tools to problems with very specific requirements and unique needs
- Adapting workflow and dataflow techniques for vastly different scales (from individual laboratory information systems all the way up to large multi-national collaborations)
- Managing the quality and provenance of information; for the scientific data itself, for the workflow specifications (and their various versions), and for data products and visualizations of the data

Another theme was that these systems tend to be built in isolation with little learning or re-use from other projects. Although many systems are currently being built, there is not an easy way for one group to learn about what is being done by other groups, unless there happens to be a person in common with both projects. Also, it is difficult for a group that is beginning to develop a workflow for a particular application to learn about the universe of tools and solutions that are available. It might be useful to have some repository of "best practices" or "experiences" for developing workflows, so that developers do not have to start from scratch each time.

A third theme was that developers of "new and exciting" solutions need to be willing to work with scientists and use their existing legacy tools and processes. Scientists become very attached to the GUIs they know how to use, to the information systems they have already spent time and resources developing, and to specialized codes (often written in FORTRAN) that they trust. Insisting that an application be ported to a new language or provide a new interface is often not feasible. Thus, much of the challenge in building these systems is to retain the components that scientists want to keep, while connecting them in new ways to facilitate better and more interesting functionality.

3. Papers and presentations

The workshop program included 10 papers, which can be roughly categorized into papers on "tools" and papers on "case studies."

Several case studies of developing requirements and systems for specific applications were presented. These case studies illustrate how widely useful workflow and data flow systems are in modern science. Scott Klasky of Oak Ridge National Laboratory discussed requirements for analyzing plasma simulation data, and focused on the need for flexibility in adapting and deploying workflows of parallel physics codes. Laura Bright of Portland State University discussed applying the "factory" metaphor to managing large numbers of data-product-generation workflows to produce as many data products as possible within an allotted time period. The key challenge is to make the best use of the "factory floor" (e.g., the available high performance computing resources) to produce as many data products as possible. Simon Cox of the University of Southampton described using the Windows Workflow Foundation product to manage wind-tunnel experiments, and discussed analyzing and visualizing data in real-time (so that problems can be detected quickly to avoid wasting an experimental run.)

Mirek Riedewald of Cornell University reported on experiences managing several different large-scale data flows. One such data flow is sky survey data from the Arecibo telescope, where the main challenge is efficiently dealing with large amounts of data. Another data flow results from the CLEO high energy particle physics experiments, which required adding provenance support to a large body of legacy code. The third data flow is a large collection of World Wide Web data (the Web Lab) for sociological studies, where the main challenge is navigating multiple large snapshots of the web.

In terms of tools, several groups are building general purpose or generalizable tools, although in many cases these tools are motivated by specific applications. Louiqa Raschid of the University of Maryland described adapting an enterprise-style mediator system, DB2 and WebSphere, for use in scientific applications. The key idea is to express the workflow as a large SQL query, and then utilize the mediation capabilities to efficiently execute and monitor the workflow. Bettina Kemme of McGill University described Exp-WF, a workflow system for managing data in laboratory-scale information systems. Her experience demonstrated the importance of keeping some legacy components (e.g., the laboratory information systems themselves), and using a paradigm of "add to, don't replace" when developing the system's functionality. Bertram Ludäscher of the University of California at Davis argued that dataflow process networks are a natural model for specifying data-intensive scientific workflows, but that control-flow and plumbing-intensive tasks lead to "messy" dataflow designs. He proposed an approach that allows one to nest state-machines (for flexible control-flow) within dataflow networks, resulting in simpler, more reusable workflows. Reusability is further enhanced by a "workflow template" mechanism. Louiqa Raschid (presenting on behalf of Zoe Lacroix from Arizona State University) described the SemanticBio system, where the workflow can be specified using high-level ontologies, separating the specification from the implementation details.

Workflow tools can provide additional value to scientists by helping them manage the provenance and quality of their data (and of the workflows themselves). Juliana Freire of the University of Utah described how scientists often spend many hours tuning and tweaking visualization workflows to provide just the right view of their data. The VisTrails system, that she and her colleagues are developing at Utah, maintains detailed information about the exploratory process - the trial-and-error steps followed to construct a set of data products. By capturing the provenance of both the derived data and the processes that generate these data, and by providing an intuitive interface for comparing the results of different workflows, VisTrails greatly simplifies the scientific discovery process. For example, it allows a scientist to manage the versions of their visualization workflows, reverting to an earlier version if necessary, comparing different visualizations side by side, and so on. Yogesh Simmhan of Indiana University described a quality model for collaborative data that allowed scientists to evaluate data on multiple axes: the metadata, the provenance, the quality of service when accessing the data, and the community evaluation of the data.

4. Moving forward

Certainly, scientific workflow and data flow systems will continue to be built, as scientists increasingly recognize their usefulness. The challenge for computer scientists is to develop tools and techniques that ease the process of creating, maintaining and executing workflows, and allow scientists to focus their energies on the science, and not on the "plumbing."

As mentioned above, building a repository or forum for sharing experiences and best practices for these systems would be a significant help to the community. One participant noted that just having people who have actually built these systems in the same room to discuss their experience is a great way to find out about new tools and avoid repeating mistakes.

In particular, sharing "tips and tricks" is important for at least two levels: the "IT" level, where computing centers are deploying tools for use by their scientists, and at the "research level," where computer scientists are developing new techniques and algorithms based on real requirements from natural scientists. One possibility for such a forum is to repeat this workshop; another possibility is to form some sort of working group that could sponsor a portal, newsletter or other forum.

Also, the need to integrate heterogeneous data, heterogeneous systems, new and legacy codes, and so on means that any tools and techniques developed in the future need to "play nicely with others" in order to have a realistic chance of adoption. Forcing scientists to use a particular programming language, operating system or data format is infeasible, given both the existing base of software and the specialized hardware requirements of various groups.

Acknowledgements

We would like to thank the authors for their compelling papers and interesting work as well as the lively discussion at the workshop. We would like to thank the program committee for providing high quality reviews, even over their holiday break. We would like to thank the ICDE organizers for providing us with a venue. We would finally like to thank Microsoft for their support of the workshop and for their commitment to help move this area forward.

The SciFlow 2006 webpage is at http://www.cc.gatech.edu/~cooperb/sciflow06/.

Papers

[1] William Y. Arms, Selcuk Aya, Manuel Calimlim, Jim Cordes, Julia Deneva, Pavel Dmitriev, Johannes Gehrke, Lawrence Gibbons, Christopher D. Jones, Valentin Kuznetsov, Dave Lifka, Mirek Riedewald, Dan Riley, Anders Ryd, Gregory J. Sharp. "Three Case Studies of Large-Scale Data Flows."

[2] Shawn Bowers, Bertram Ludaescher, Anne H.H. Ngu, Terence Critchlow. "Enabling Scientific Workflow Reuse through Structured Composition of Dataflow and Control-Flow."

[3] Laura Bright, David Maier, Bill Howe. "Managing the Forecast Factory."

[4] Steven P. Callahan, Juliana Freire, Emanuele Santos, Carlos E. Scheidegger, Claudio T. Silva, Huy T. Vo. "Managing the Evolution of Dataflows with VisTrails."

[5] Barbara A. Eckman, Terry Gaasterland, Zoe Lacroix, Louiqa Raschid, Ben Snyder, Maria Esther Vidal. "Implementing a Bioinformatics Pipeline (BIP) on a Mediator Platform: Comparing Cost and Quality of Alternate Choices."

[6] Brian Gabor, Bettina Kemme. Exp-WF: Workflow Support for Laboratory Information Systems.

[7] Scott A. Klasky, Bertram Ludaescher, Manish Parashar. "The Center For Plasma Edge Simulation Workflow Requirements."

[8] Herve Menager and Zoe Lacroix. "A Workflow Engine for the Execution of Scientific Protocols."

[9] A. Paventhan, Kenji Takeda, Simon J. Cox, Denis A. Nicole. "Leveraging Windows Workflow Foundation for Scientific Workflows in Wind Tunnel Applications."

[10] Yogesh L. Simmhan, Beth Plale, Dennis Gannon. "Towards a Quality Model for Effective Data Selection in Collaboratories."