IEEE Computer, vol. 40, no. 12, pp. 24-32, December, 2007.

Examining the Challenges of Scientific Workflows

Yolanda Gil[1], Ewa Deelman[1], Mark Ellisman[2], Thomas Fahringer[3], Geoffrey Fox[4], Dennis Gannon[4], Carole Goble[5], Miron Livny[6], Luc Moreau[7], Jim Myers[8]
[1]USC Information Sciences, [2]University of California San Diego, [3]Innsbruck University, [4]Indiana University, [5]Manchester University, [6]University of Wisconsin Madison, [7]University of Southampton, [8] National Center for Supercomputing Applications
gil@isi.edu, deelman@isi.edu, mellisman@ucsd.edu, tf@dps.uibk.ac.at, gcf@grids.ucs.indiana.edu, gannon@cs.indiana.edu, carole.goble@manchester.ac.uk, miron@cs.wisc.edu , L.Moreau@ecs.soton.ac.uk, jimmyers@ncsa.uiuc.edu

Abstract

Workflows have recently emerged as a paradigm for representing and managing complex distributed scientific computations and therefore accelerate the pace of scientific progress. A recent workshop on the Challenges of Scientific Workflows, sponsored by the National Science Foundation and held on May 1-2, 2006, brought together domain scientists, computer scientists, and social scientists to discuss requirements of future scientific applications and the challenges that they present to current workflow technologies. This paper reports on the discussions and recommendations of the workshop, the full report can be found at http://www.isi.edu/nsf-workflows06.

1. Introduction

Significant scientific advances are increasingly achieved through complex sets of computations and data analyses. These computations may comprise thousands of steps, where each step may integrate diverse models and data sources developed by different groups. The applications and data may be also distributed in the execution environment. The assembly and management of such complex distributed computations present many challenges, and increasingly ambitious scientific inquiry is continuously pushing the limits of current technology.

Workflows have recently emerged as a paradigm for representing and managing complex distributed scientific computations and therefore accelerate the pace of scientific progress [1,2,3,4,5,6]. Scientific workflows capture the individual data transformations and analysis steps as well as the mechanisms to carry them out in a distributed environment. Each step in the workflow specifies a process or computation to be executed (e.g., a software program to be executed, a web service to be invoked). The steps are linked according to the data flow and dependencies among them. The representation of these computational workflows contain many details required to carry out each analysis step, including the use of specific execution and storage resources in distributed environments, Workflow systems can exploit these explicit representations of the complex computational processes to manage their lifecycle and to automate their execution. Workflows can capture complex analysis processes at various levels of abstraction, and also provide the provenance information necessary for scientific reproducibility, result publication, and result sharing among collaborators. By providing formalism and by supporting automation, workflows have the potential to accelerate and transform the scientific analysis process. Existing workflow systems have been demonstrated in a variety of scientific applications, were workflows composed of thousands of components processed large-distributed data sets on high-end computing resources. Some workflow systems have been deployed for routine use in scientific collaboratories. Figure 1 shows an image created by an astronomy application, Montage [7]. Montage uses workflow technologies [8] to generate science-grade mosaics of the sky. Such mosaics were recently used to verify a bar in the M31 galaxy [9]. Although there have been hints of a bar in M31 from optical data, none of the analyses were convincing because the effects of interstellar extinction at optical wavelengths were severe. However, the universe is much more transparent in the infrared, and this enabled astronomers to overcome the effects of interstellar extinction. There was one more problem: the variable background in the infrared images hid the structure of

the galaxy. By using Montage, which is able to rectify the backgrounds to a common level, astronomers were finally able to see the bar structure.

Much research is underway to address issues of creation, reuse, provenance tracking, performance optimization, and reliability. However, to fully realize the promise of workflow technologies, many additional requirements and challenges must be met. Scientific applications are driving workflow systems to examine issues such as supporting dynamic event-driven analyses, handling streaming data, accommodating interaction with users, intelligent assistance and collaborative support for workflow design, and enabling result sharing across collaborations. As a result, a more comprehensive treatment of workflows is needed to meet long-term requirements of scientific applications.

To examine the nature of these challenges and to consider what steps should be taken to address them, a Workshop on the Challenges of Scientific Workflows was held at the National Science Foundation on May 1-2, 2006. The meeting brought together domain scientists, computer scientists, and social scientists to discuss requirements of future scientific applications and the challenges that they present to current workflow technologies.

This report summarizes the discussions and recommendations of the workshop. The workshop discussions focused on four main topics, summarized in the following four sections. The final section of the paper provides conclusions and recommendations of the workshop attendees. An overview and pointers to the state of the art in the area of scientific workflows is provided in the workshop report [1] (available at http://www.isi.edu/nsf-workflows06), including citations and references to relevant work.

2. Discussion Topic I: Application Requirements

A key motivating question posed by domain scientists participating in the workshop was:

> ***Given the exponential growth in computing, sensors, data storage, network, and other performance elements, why is the growth of scientific data analysis and understanding not proportional?***

There was a broad consensus in the group that in the scientific community there is a perceived importance of workflows in accelerating the pace of scientific discoveries. Today, complex scientific analyses increasingly require tremendous amounts of human effort and manual coordination. Data is growing exponentially, but the number of scientists is roughly constant. Thus researchers need exponentially more effective tools to aid in their work, if they are not to be inundated in data and associated tasks. Workflow environments that support and improve the scientific process at all levels are crucial if we are to sustain the current rapid growth rate in data and processing.

The ability to combine distributed data, computation, models, and instruments at unprecedented scales can enable transformative research. The analysis of large amounts of widely distributed data is becoming commonplace. These data, and the experimental apparatus or simulation systems that produce them, typically do not belong to individuals but rather to collaborations. Within these collaborations, various individuals are responsible for different aspects of data acquisition, processing, and analysis, and in which publications are often generated by entire projects. Such environments demand tools that can orchestrate the steps of scientific discovery and bridge between the differing expertise of the members of the collaboration.

Many disciplines are benefiting from the use of workflow management systems to automate such computational activities. Examples of such disciplines include astronomy, biology, chemistry, environmental science, engineering, geosciences, medicine, physics, and social sciences. The workshop report provides pointers and detailed examples to this work.

An important application requirement identified by workshop participants is **reproducibility of scientific analyses and processes**. This requirement is at the core of the scientific method, in that it enables scientists to evaluate the validity of each other's hypothesis and provides the basis for establishing known truths. Reproducibility requires rich *provenance* information, so that researchers can repeat techniques and analysis methods to obtain scientifically similar results. Today, reproducibility is virtually impossible for complex scientific applications. First, because so many scientists are involved, the provenance records are highly fragmented, and in practice they are reflected in a variety of elements including emails, Wiki entries, database queries, journal references, codes (including compiler options), and others. All this information, often stored in a variety of locations and in a variety of forms, needs to be appropriately indexed and made available for referencing. Without tracking and integrating these crucial bits of information together with the analysis results, reproducibility can be largely impractical, and more likely impossible, for

many important discoveries involving complex computations.

In order to support reproducibility, workflow management systems must **capture and generate provenance information as a critical part of the workflow-generated data.** Workflow management systems must also consume the provenance information associated with input data, and associate that information with the resulting data products. Provenance must be associated and stored with the new data products and contain enough details to enable reproducibility. Another important requirement is for interoperable, persistent repositories of data and analysis definitions, with linkage to open data and publications, as well as to the algorithms and applications used to transform the data. Existing data repositories must be complemented with provenance and metadata repositories that enable the discovery of the workflows and application components that were used to create the data. An important concern for scientists in these highly collaborative endeavors is credit assignment and recognition of individual contributions.

The environments provided should also be flexible in terms of **supporting both common analyses performed by many as well as unique individual analyses**. Routine analyses based on common cases should be easy to set up and execute. At the same time, individual scientists should be able to steer the system to conduct unique analyses and to create novel workflows with previously unseen combinations and configurations of models.

From an operational perspective, there is a need to provide **solutions that are secure, reliable, and scalable**. Scientists need to be able to trust that their input and output data are secure and free from inappropriate data access or malicious manipulation. Trust and reputation systems for data providers must be incorporated into current infrastructure. Tools need to be scalable in order to support large and complex analyses, TeraByte and greater size data sets, and large scientific communities.

An important concern is how to **address the inevitable heterogeneities and inconsistencies that arise when information comes from different sources and communities**. Mechanisms for curating, validating, translating, and integrating data are needed in order for scientific information to be shared in meaningful and truly integrative ways.

Finally, scientists need **easy to use tools that provide intelligent assistance for such complex workflow capabilities**. Automation of low-level operational aspects of workflows is a key requirement. Interaction modalities that hide unnecessary complexities and speak the scientist's language will be crucial to success. Guidance to users will be useful to encourage the best scientific practices.

3. Discussion Topic II: Data and Workflow Descriptions

A key issue addressed by this discussion group was:

> ***Given the broad practice and many benefits of sharing instruments, data, computing, networking, and many other science products and resources, why are scientific computations and processes not widely captured and shared as well?***

Scientists have always relied on technology to share information about experiments, from pen and paper to digital cameras, email, the Web, and computer software. Workflow description and execution capabilities offer a new way of sharing and managing information: one in which full processes can be captured electronically and shared for future reference and reuse. This new way of sharing information—agreeing on semantics of processes themselves and the infrastructure to support their execution—continues the historic push for making representations explicit and actionable, and reducing the barriers to coordination. **Scientists should be encouraged to bring workflow representations to their practices and share the descriptions of their scientific analyses and computations in ways that are as formal and as explicit as possible.** However, there are no commonly accepted and sufficiently rich representations in the scientific community. Thus, more research in this area is needed.

Workflow representations need to **accommodate scientific process descriptions at multiple levels**. For instance, domain scientists may want a sophisticated graphical interface for composing relatively high-level scientific or mathematical steps, whereas computer scientists may be more concerned with the use of a workflow language, and with the detailed specifications of data movement and job execution steps. To link between these views and to provide needed capabilities, workflow representations must include rich descriptions that span abstraction levels, and must include models of how to map between them.. Further, to support the end-to-end description of multidisciplinary, community-scale research,

definitions of workflow and provenance must be broad enough to describe workflows-of-workflows that are linked through reference data, models backed by validation workflows, the scientific literature, and manual processes in general. Other important dimensions of abstraction are experiment-critical vs. non-experiment-critical representations, where the former refers to scientific issues and the latter is more concerned with operational matters. A workflow system should support both sets of concerns.

Rich information about analysis processes needs to be incorporated in workflow representations to **support workflow discovery, creation, merging, and execution**. These activities will become a natural way to conduct experiments and share scientific methodology within and across scientific communities.

Workflow representations need to **support, wherever possible, automation of the workflow creation and management processes**. This capability will require rich semantic representations of requirements and constraints on workflow models and components. With semantic descriptions of the data format and type requirements of a component, it is possible to incorporate automated reasoning and planning capabilities that could automatically add data conversion and transformation steps. Similarly, with rich descriptions of the execution requirements of each workflow component, automated resource selection and dynamic optimizations would be possible.

A challenge for the computer science community is to be able to **manipulate the multiple levels of workflow abstraction simultaneously and to manipulate them individually**. For instance, several distinguishable levels of process abstraction were considered useful in the breakout group: scientific, engineering, and instance. Another classification distinguished among data description, functional behavior specification, non-functional aspects, and execution/run-time aspects. A capability of "workflow abstraction" would allow scientists to identify what level(s) of description are useful to share in their workflows, and package such a description as a self-contained sharable object, which can then be refined and instantiated by other scientists. Refinement and abstraction capabilities are needed for all first-class entities that have to be manipulated by workflow systems: workflow scripts (regarded as specifications of future execution), provenance logs (descriptions of process and data history), data, and metadata. There is relevant work in related fields of computer science, such as refinement calculi, model-driven architectures, and semantic modeling, but these techniques have not been applied widely to scientific workflows, which are potentially large scale, may involve multiple technologies, and have to operate on heterogeneous systems. We also note that sophistication of descriptions needed is dependent on the workflow capabilities needed. For example, a workflow that adapts dynamically to changes in environment or data values requires formal and comprehensive descriptions so that a machine can make a decision on adaptation. Even for a human to make choices related to making changes to a workflow would require access to a broad variety of descriptions.

Another important research issue is **whether scientific workflows can or even should build on existing workflow technologies**, or whether they require fundamentally new approaches. Workflows have been used for decades to represent and manage business processes. There are emerging standards for workflow representations as well as associated software (some of commercial quality) to manage workflows. Understanding the differences between scientific workflows and practices and those used in business could yield useful insights. On the one hand, scientific and business workflows are not obviously distinguishable, since both may share common important characteristics. Indeed, in the literature, we find examples of workflows in both domains that are data intensive, highly parallel, etc. On the other hand, scientific research requires flexible design and exploration capabilities that appear to depart significantly from the more prescriptive use of workflow in business; **workflows in science are a means to support detailed scientific discourse as well as a way to enable repeatable processes**.

Another distinctive issue of scientific workflows is the variety and heterogeneity of data within a single workflow. For example, scientific workflow may involve numeric and experimental data in proprietary formats (such as those used for raw data produced by the scientific instruments involved in a process), followed by processed data resulting in description related to scientific element (e.g., molecule or biochemistry descriptions), leading to textual, semi-structured, and structured data, and formats used for visual representation. To clarify the research issues in developing scientific workflow capabilities, the community needs to identify where there are real differences between scientific and business activities, beyond domain-specific matters. An important concern is to balance the desire for sharing workflow information against the dangers of premature standardization efforts that may constrain future requirements and capabilities. In this respect, it will be crucial to encourage computer scientists and domain scientists to collaborate closely in developing more

workflow-based applications and to discuss representation requirements for future workflows.

In the discussion, it was recognized that most scientific activity consists of exploration of variants and experimentation with alternative settings, which would involve **modifying workflows to understand their effects and how to explain those effects**. Hence, an important challenge in science is representation of workflow variants, which aims at understanding the impact that a change has on the resulting data products as an aid to scientific discourse. As part of managing change, version control becomes important. The challenge of evolving workflows is further compounded by the need to validate data products and to disseminate and share experiments and data within the scientific community. Hence, traceability and sharing are key requirements of scientific workflows.

While acknowledging that the sharing of representations is important to the scientific process, the group recognized that **multiple collaboration and sharing practices must be accommodated**. In some cases, it is suitable to share workflows, but not data. In other cases, scientists want to share an abstract description of the scientific protocol, without actually communicating details, parameters and configurations, which are their private expertise. In other situations, it is a description of a specific previous execution (provenance) that is desirable, with or without providing execution details.

4. Discussion Topic III: Dynamic Workflows and User Steering

The participants in this discussion were tasked with examining issues related to the dynamic nature of the scientific analysis and focused on the following question:

> ***How can workflows support the exploratory nature of science and the dynamic processes involved in the scientific analysis?***

Given that the experimental context of the user is in flux (as the scientific discovery process evolves) and the distributed infrastructure that the workflows operate over is in flux (as networks, platforms and other resources come and go), the notion of static workflows is an odd one. The vision of supporting dynamic, adaptive and user-steered workflows is to enable and accelerate distributed and collaborative scientific methodology via rapid reuse and exploration and continuous adaptation and improvement. Reproducibility becomes ever more elusive in this kind of setting. The challenge is to develop mechanisms to create, manage, and capture dynamic workflows so that reproducibility of significant results is possible.

Scientific practice will routinely give rise to workflows that are dynamic where the decisions they make about which steps to take next are based on the latest available information. A workflow may need to be dynamically designed in the sense of looking at the results of the initial steps before a decision can be made about how to carry out later analysis steps. For example, by examining the results of some initial pre-processing of an image subsequent steps may be needed to look at specific areas identified by that pre-processing. A dynamic workflow could also be one where the basic structure or semantics or the workflow changes because of some external event. For example, in severe storm prediction, data analysis agents may examine radar data searching for specific patterns. Depending upon the specific pattern of events, different branches of a storm prediction workflow may be enacted which may require that significant computational resource be made available on-demand. Should the storm intensify or should resource availability change, the workflow must adapt. Some experimental regimes may draw on workflows that are heuristic or that employ untried activities, and thus these workflows may breakdown or fail during their execution, thereby necessitating fault diagnosis and repair. Another scenario which includes dynamic workflows is where two workflows could affect each other, for example by sharing results. They can be classified as dynamic as they respond to events arising in each other's execution. Finally, some scientific endeavors are large-scale. They involve large teams of scientists and technicians, and engage in experimental methods or procedures that take long times to complete and require human intervention and dynamic steering throughout the process. For example, the study of deep-space phenomena in astrophysical studies may require the use and coordination of multiple observation devices operating in different spaces, capturing data at different frequencies or modalities, and the resulting data will need to be cleaned and aligned for proper interpretation. Any step in such scientific inquiry may be subject to both the exigencies of sensor operation, weather or spatial occlusions during scheduled observation periods, and other delays, not to mention reactive adjustments to later stage observations arising from preliminary discoveries in earlier observational steps.

The **management of dynamic workflows is complex due to their evolution and lifecycle**. There is no beginning or end to the lifecycle process of a workflow – scientists can start at any point and flow through the figure in any direction. They might build

or assemble a workflow, refine one that has previously been published to a shared repository, run their design, evolve it, run it again, share fragments of it as they go along, find other fragments they need, run it a few more times, and learn from the protocol they are developing. They might settle on the workflow and run it many times, learning from the results produced, or maybe they run it just once, because that is all they need. While running, the workflows could adapt to external events and user steering. The results of the whole activity feed into the next phases of investigation. The user is ultimately at the centre, interacting with the workflows and interpreting the outcomes.

Supporting scientists in complex exploratory processes involving dynamic workflows is an important challenge. A human-centered decision support system that accommodates the information needs of a scientist tracking and understanding such complex processes will need to be designed. Appropriate user interfaces that enable scientists to browse/traverse, query, re-capitulate, and understand this information will be needed. Simplifying the exploratory process also requires novel and scalable means for scientists to manipulate the workflows, explore slices of the parameter space, and compare the results of different configurations. Easily assembling workflows, finding services and adapting previous workflows is key.

An interesting direction for future research explores the question of how to **improve, redesign, or optimize workflows through data mining of workflow lifecycle histories** to learn successful (and unsuccessful) workflow patterns and designs and assist users to follow (or avoid) them. One kind of pattern can be extracted from successful execution trails. This information can be used to build recommendation systems. For example, if a model M is added, the system could suggest additional models that other people often use together with M in a workflow or suggest values commonly used for the parameters in the model. Another kind of pattern could be extracted from unsuccessful trails. These can, for example, help identify incompatible parameter settings, unreliable servers or services, gross inefficiencies in resource usage, etc. Workflow patterns can subsequently be analyzed, re-enacted (reproduced), and validated in order to facilitate their reuse, continuous improvement, and redeployment into new locations or settings.

5. Discussion Topic IV: System-level Workflow Management

A key issue addressed by this discussion group was:

Given the continuous evolution of infrastructure and associated technology, how can reproducibility of computational analyses be ensured over a long period of time?

A key challenge in scientific workflows is **ensuring engineering reproducibility to enable the re-execution of analyses, and the replication of results**. Scientific reproducibility implies that someone can follow the general methodology, relying on the same initial data, and obtain equivalent results. Engineering reproducibility requires more knowledge of the data manipulations, of the actual software and execution environment (hardware, specific libraries), etc., so that the results can be replicated bit-by-bit. The former capability is needed when researchers want to validate each other's hypotheses, whereas the latter is beneficial when unusual results or errors are found and their source needs to be traced and understood. The information needed to support both types of reproducibility is challenging to capture. When supporting scientific reproducibility, a high-level, yet meaningful, description of the process needs to be provided. Engineering reproducibility also necessitates low-level information such as what compiler flags were used to compile a particular code and the details of the execution environment and computer architecture.

An important challenge will be to **provide a stable view on the system in spite of continuous changes in technology and platforms at the system level**. The underlying execution system must be designed so that it provides a stable environment for the software layers managing the high-level scientific process. It must be possible to re-execute workflows many years later and obtain the same results. This requirement poses challenges in terms of creating a stable layer of abstraction over a rapidly evolving infrastructure while providing the flexibility needed to address evolving requirements and applications and to support new capabilities. In order to provide consistent and efficient access to resources, resource management must consider both physical resources (e.g., computers, networks, data servers) and logical resources (e.g., data repositories, programs, application components, workflows). Both should be exposed through uniform interfaces. By enhancing resource descriptions with semantic annotations, the provisioning, provenance, configuration and deployment of new resources can be organized more easily and possibly even automated. Extending current information services with meaningful semantic description of resources should

enable semi-automatic discovery, brokering, and negotiation. Human interaction should be minimized through dynamic configuration and lifecycle management of resources. Some efforts have been made to provide semi-automatic discovery and brokering of physical resources and management of software components that may become part of scientific workflow environments. However, there is still much opportunity for improvements, since most existing systems require manual or semi-manual deployment of software components and force application builders to hardcode software component locations on specific resources into their workflows. Additionally, currently available information services are not well adapted to store complete description of software components, forcing the application builder to use only (name, location)-style information about available services and resources. As a consequence these applications are sensitive to dynamic changes in the resource infrastructure, and often fail during execution due to avoidable failures.

Workflow end users frequently want to be able to specify quality of service requirements. These requirements then should be guaranteed—or at least maintained on a best effort basis—by the underlying runtime environment. However, current systems are mostly restricted to best effort optimizations for time-based criteria such as reducing overall execution time or maximizing bandwidth. Several problems must be addressed to overcome current limitations. First, quality of service parameters need to be extended beyond time-based criteria to cover other important aspects of workflow behavior such as responsiveness, fault tolerance, security, and costs. This effort will require collaborative work on the definition of quality of service parameters that can be widely accepted among scientists, so as to provide a basis for interoperable workflow environments or services. Current optimization and planning approaches may have to be radically changed to cope with multi-criteria optimization or planning. Many systems exist for single and some for bi-criteria optimization, but hardly any systems tackle multi-criteria optimization problems. There is no ready-to-use methodology that can deal with this problem in an efficient and effective way; thus, there are many opportunities for research. In developing runtime environment support for quality of service, reservation mechanisms will be an important tool. Both immediate and advance reservations can make the dynamic behavior of infrastructures more predictable, an important prerequisite to guarantee quality of service such as responsiveness and dependability. Moreover, advance reservation can also simplify the scheduling of workflow tasks to resources. However, reservations also introduce challenges relating to policy (who gets to make reservations), fragility (in contrast to a best effort resource, reservable resources may suddenly become unavailable due to a reservation), and efficiency of resource utilization. In providing reservation mechanisms, we should address not only physical resources but also logical resources such as Web services, licenses, and executables. It should be the task of resource management systems to guarantee reservation of physical resources on which logical resources are executed or processed.

Challenging **issues of scale arise in workflow execution**, and these issues will increasingly require advances over the current state of the art. These issues occur in multiple dimensions. First, we see individual workflows becoming increasingly large in many disciplines, as (for example) the quantities of data operated on become larger. As workflows scale from 1,000 to 10,000 and perhaps 1,000,000 tasks or more, new techniques may be needed to represent sets of tasks, manage those tasks, dispatch tasks efficiently to resources, monitor task execution, detect and deal with failures, and so on. A second important scaling dimension is the number of workflows. Particularly in large communities, many users may be submitting many workflows at once. If these workflows compete for resources or otherwise interact, then appropriate supporting mechanisms are needed in the runtime environment to arbitrating among competing demands. A third scaling dimension concerns the number of resources involved. Ultimately, we can imagine tasks running on millions of data and computing resources (indeed, some systems such as SETI@home already do operate at that scale). A fourth scaling dimension concerns the number of participants. In a simple case, a single user prepares and submits a workflow. In a more complex case, many participants may be involved in defining the workflow, contributing relevant data, managing is execution, and interpreting results.

New infrastructure services to support workflow management must be provided. Some of these services are analogous to existing data management and information services: for example, workflow repositories and workflow registries. Other more novel services will be concerned with workflows as active processes, and the management of their execution state.

An important issue to address is the **perceived tension between research challenges of scientific workflows and the constraints imposed by existing production-quality infrastructure**. Shared

infrastructures such as the TeraGrid and NMI[1] provide widely used and well-tested capabilities to build on. These system-level infrastructure layers are designed to be production quality, but out of necessity have not been designed to address specific requirements of scientific workflows. Rather, they aim to meet the needs of a broader research community. It is unlikely that commitments can be made at this point by selecting particular architectures or implementations at the workflow layers of shared cyberinfrastructure. Alternative architectures must be explored to understand design tradeoffs in different contexts: for example, workflows designed and tested on a person's desktop that are then run with larger data in a cluster, workflows to handle streaming data, event-driven workflow management engines, and architectures centered on interactivity. At the same time, these architectures could be designed to be interoperable and compatible, where feasible, with some overall end-to-end, multi-level framework. Follow-on discussions and workshops to understand and address these issues will be extremely beneficial.

6. Concluding Remarks and Recommendations

Workflows provide a formal specification of the scientific analysis process from the data collection, through analysis to the data publication. Workflows can be viewed as recipes for cyberinfrastructure computations, providing a representation to describe the end-to-end processes involved in carrying out heterogeneous interdependent distributed computations.

Once this process is captured in declarative workflow structures, workflow management tools could **accelerate the rate of scientific progress** by supporting scientists in creating, merging, executing, and re-using these processes. By assisting scientists in reusing well-known and common practices for analyses, complex computations will become a daily commodity for scientific discovery. By coaching scientists to conduct experiments in neighboring disciplines, cross-disciplinary scientific analyses will become commonplace.

Scientists view **workflows as key enablers for reproducibility of experiments involving large-scope computations.** Reproducibility is engrained in the scientific method, and there is a concern that without this ability there will be a rejection of cyberinfrastructure as a legitimate means to conduct scientific experiments. To enable reproducibility, workflow management systems are needed to capture the end-to-end process at all levels of abstraction, from the science domain level down to the system level. This information is generally termed as **provenance** and is key to reproducibility. Representing scientific processes with enough fidelity and flexibility will be a key challenge for the research community. Recognizing that science has an exploratory and evolutionary nature, workflows need to support dynamic and interactive behavior. Thus workflow systems need to become more dynamic and amenable to steering by users and be more responsive to changes in the environment.

7. Summary of Recommendations

The following recommendations were made by the workshop participants:

- **Support basic research in computer science to create a science of workflows**. Although existing systems are addressing important issues such as workflow creation, planning, and execution, more comprehensive research is needed to provide easy-to-use workflow construction tools, develop sophisticated automation tools, provide robust workflow execution, manage complex dynamic workflows, etc. There are many open research issues to be resolved in computer science proper that will enable significant progress in the research agenda of scientific workflows.
- **Make explicit workflow representations that capture scientific analysis processes at all levels the norm when performing complex distributed scientific computations.** We need workflow representations at different levels of abstraction, so that we can represent workflows at different levels of refinement, from abstract application-level definition down to operational, system-specific description. These workflow representations can become a starting point for defining common representations that can be interpreted by a variety of workflow systems.
- **Integrate workflow representations with other forms of scientific record.** Data created through workflows should include representations of those workflows as metadata. Articles in scientific publications should include not only textual descriptions of the processes utilized, but also

[1] www.teragrid.org, www.nsf-middleware.org.

formal descriptions specified as workflows. Laboratory notebooks and invention records should be annotated with workflows and the rationale for their design and final configuration.

- **Support and encourage cross-disciplinary projects involving relevant areas of computer science as well as domain sciences with distinct requirements and challenges.** Cross-disciplinary projects between computer scientists and application scientists are needed to ensure that research efforts are directed towards areas where they can have a significant impact. Other disciplines, such as social sciences and cognitive science should also be engaged to meet the stated challenges.
- **Provide long-term, stable (five or more years) collaborations and programs.** Based on the experiences of the NSF ITR program and the UK e-Science program, the greatest successes were obtained in collaborations that were funded for five years or more, so that collaborations had time to mature and obtain significant results.
- **Define a roadmap to advance the research agenda of scientific workflows while building on existing cyberinfrastructure.** Significant investment in cyberinfrastructure, has resulted in production quality services for data management, high-end and large-scale computation, resource sharing, and distributed computing. It will be important to articulate anticipated requirements to support scientific workflows in the coming years, and develop a roadmap for how the current infrastructure can evolve to accommodate the challenging research agenda that lies ahead. A follow-up workshop on this topic in the near future would be highly beneficial.
- **Coordinate between existing and new projects on workflow systems and interoperation frameworks for workflow tools**. Many current projects have evolved in isolation, working with non-intersecting scientific communities. Capturing best practices will enable a better understanding of the existing capabilities. It will be beneficial to consider the development of a common framework, so that various workflow tools can be integrated and interchanged with others. Scientists will then be able to concentrate on the science rather than have to worry about the particulars of different workflow systems.
- **Hold follow-up, cross-cutting workshops and meetings**. More workshops are needed, to bring together scientists from various domains. Encourage discussion between sub-disciplines of computer science, to and bring in human factors and collaboration considerations to workflow management systems.

In summary, **workflows should become first-class entities in the cyberinfrastructure architecture.** For domain scientists, they are important because workflows document and manage the increasingly complex processes involved in exploration and discovery through computation. For computer scientists, workflows provide a formal and declarative representation of complex distributed computations that must be managed efficiently through their lifecycle from assembly, to execution, to sharing.

8. Acknowledgements

This workshop was sponsored by the National Science Foundation under grant # IIS-0629361. We would like to thank Maria Zemankova, Program Manager of the Information and Intelligent Systems Division, for supporting the workshop and contributing to the discussions. We would also like to thank all the workshop attendees for their contributions: Mark Ackerman, Ilkay Altintas, Roger Barga, Francisco Curbera, Constantinos Evangelinos, Juliana Freire, Ian Foster, Alexander Gray, Jeffrey Grethe, Jim Hendler, Carl Kesselman, Craig Knoblock, Chuck Koelbel, Karen Myers, Walt Scacchi, Ashish Sharma, Amit Sheth, Alex Szalay, and Gregor Von Laszewski. The authors would also like to thank Bruce Berriman for his input and discussions.

9. References

[1] Deelman, E. and Gil, Y. (Eds.) *Final Report of the NSF Workshop on Challenges of Scientific Workflows*, National Science Foundation, Arlington, VA, May 1-2, 2006. Available at http://www.isi.edu/nsf-workflows06, also available from NSF at the Division of Information and Intelligent Systems (IIS) site http://www.nsf.gov/events/event_summ.jsp?cntn_id=108411&org=IIS and at the Office of Cyberinfrastructure (OCI) site http://www.nsf.gov/od/oci/reports.jsp.

[2] Deelman, E. and Taylor, I. (Eds.) *Journal of Grid Computing*, Special Issue on Scientific Workflows, Volume 3, Number 3-4, September 2005.

[3] Deelman, E., Zhao, Z., and Belloum, A, (Eds.) *Scientific Programming Journal*, Special Issue on Workflows to Support Large-Scale Science. 2006.

[4] Fox, G. and Gannon, D. (Eds.) *Concurrency and Computation: Practice and Experience*, Special Issue on Workflow in Grid Systems. Volume 18, Issue 10, August 2006.

[5] Ludaescher, B. and Goble, C. (Eds.) *SIGMOD Record*, Special Issue on Scientific Workflows, Volume 34, Number 3, September 2005.

[6] Taylor, I.J.; Deelman, E.; Gannon, D.B.; Shields, M. (Eds.) *Workflows for e-Science: Scientific Workflows for Grids*, Springer Verlag, 2006.

[7] Berriman, G. B. et al., "Montage: A Grid Enabled Engine for Delivering Custom Science-Grade Mosaics On Demand," in SPIE Conference 5487: Astronomical Telescopes, 2004.

[8] Deelman, E., et al., "Pegasus: a Framework for Mapping Complex Scientific Workflows onto Distributed Systems," Scientific Programming Journal, vol. 13, pp. 219-237, 2005

[9] Beaton, R. L, et al., Unveiling the Boxy Bulge and Bar of the Andromeda Spiral Galaxy, Astrophysical Journal Letters (in submission).

[10] Explanatory Supplement to the 2MASS All Sky Data Release and Extended Mission Products http://www.ipac.caltech.edu/2mass/releases/allsky/doc/explsup.html