NSF Workshop on the Challenges of Scientific Workflows

Ewa Deelman and Yolanda Gil (co-chairs)

USC Information Sciences Institute

http://www.isi.edu/nsf-workflows06

May 1-2, 2006
National Science Foundation, Arlington VA
Overview and Summary

- Current CyberInfrastructure is an enabler of a significant paradigm change in science
  - Distributed interdisciplinary data deluged scientific methodology as an end (instrument, conjecture) to end (paper, Nobel prize) process is leading to a Transformative approach
  - Reproducibility, key to scientific practice, is threatened
    - Process is increasingly complex and highly distributed

- Workflows are emerging as a paradigm for process-model driven science that captures the analysis itself

- Workflows need to be first class citizens in scientific CyberInfrastructure
  - Enable reproducibility
  - Accelerate scientific progress by automating processes

- Interdisciplinary and intradisciplinary research challenges
Science under Current CyberInfrastructure

- Science is undergoing a significant paradigm change
  - Entire communities are collaborating and pursuing joint goals (NEES, SCEC, NVO, GriPhyN, BIRN)
  - Instruments, hardware, software, and other resources shared (TeraGrid, OSG, NMI)
  - Data shared and processed at large scales

- In these shared environments, workflows are emerging as a useful paradigm to:
  - Represent complex analyses
  - Manage computation
  - Capture data provenance
  - Declaratively represent analysis processes

- Scientific collaborations using workflow paradigm for analysis (SCEC, NVO, LIGO, SEEK, myGrid, others)
Computer Science Perspective

- Workflow approaches are exploring specific aspects of the problem:
  - Creation, reuse, provenance, performance, reliability

- New requirements are emerging
  - Streaming data, from batch to interactive steering, event-driven analysis, collaborative design of workflows

- Need to develop a science of workflows
  - A more comprehensive treatment of workflow lifecycle
  - Understand long-term requirements from science applications
  - Workflows as first-class citizens in CyberInfrastructure
Science Perspective

Need a more comprehensive treatment and use of workflows to support and record new scientific methodologies

- **Reproducibility** is core to scientific method and requires rich provenance, interoperable persistent repositories with linkage of open data and publication as well as distributed simulations, data analysis and new algorithms.

- **Distributed science methodology** captures and publishes all steps (a rich cloud of resources including emails, Wikis as new electronic log books as well as databases, compiler options …) in scientific process (data analysis) in a fashion that allows process to be reproducible; need to be able to electronically reference steps in process.
  - Traditional workflow like BPEL/Kepler/Pegasus/Taverna only describes a part of this

- Multiple **collaborative heterogeneous interdisciplinary** approaches to all aspects of the distributed science methodology inevitable; need research on integration of this diversity
Scientific Workflows

- Emerging paradigm for large-scale and large-scope scientific inquiry
  - Large-scope science integrates diverse models, phenomena, disciplines

- Workflows provide a formalization of the scientific analysis process
  - analysis routines need to be executed, the data flow amongst them, and relevant execution details

- Workflows provide a systematic way to capture scientific methodology and provide provenance information for their results

- Collaboratively designed, assembled, validated, analyzed

- Should be shared just like today data collections and compute resources are shared among communities
Increasing Requirements for Workflow Systems

Domain scientists are becoming ever ambitious, so are computer scientists

- Moving from batch executions to interactive, dynamic workflows
- Assistance and automation in workflow assembly
- Provenance in terms of semantic description of large volumes of data
- Comprehensive paradigm for scientific processes: data collection, processing and data mining, discovery
  - Repeatability
  - Reproducibility
- Workflows as collaborative artifacts
- Efficient and reliable large-scale execution
Goals of the Workshop

- Explore workflow challenges from a variety of perspectives
  - Grid computing
  - Web services and network infrastructure
  - Semantic technologies
  - Organization processes
  - Process planning and control
  - Application domains
  - Information Integration
  - Scientific data mining and discovery

- Strategy and recommendations to the community to meet those challenges
Attendees

- Mark Ackerman, University of Michigan
- Guy Almes, NSF OCI
- Ilkay Altintas, SDSC
- Roger Barga, Microsoft
- Francisco Curbera, IBM
- Ewa Deelman, USC Information Sciences Institute
- Mark Ellisman, UC San Diego
- Constantinos Evangelinos, MIT
- Thomas Fahringer, University of Innsbruck
- Juliana Freire, University of Utah
- Ian Foster, University of Chicago & ANL
- Geoffrey Fox, Indiana University
- Dennis Gannon, Indiana University
- Yolanda Gil, USC Information Sciences Institute
- Carole Goble, University of Manchester
- Alexander Gray, Georgia Tech
- Jeffrey Grethe, UC San Diego
- Jim Hendler, University of Maryland
- Carl Kesselman, USC Information Sciences Institute
- Craig Knoblock, USC Information Sciences Institute
- Chuck Koelbel, Rice University
- Miron Livny, University of Wisconsin
- Luc Moreau, University of Southampton
- Jim Myers, NCSA
- Karen Myers, SRI International
- Walt Scacchi, University of California Irvine
- Ed Seidel, LSU
- Ashish Sharma, Ohio State University
- Amit Sheth, University of Georgia
- Alex Szalay, John Hopkins University Physics, Astronomy
- Gregor Von Laszewski, ANL
- Maria Zemankova, NSF IIS (PM for this workshop)
Workshop products

- Wiki repository [http://vtcpc.isi.edu/wiki](http://vtcpc.isi.edu/wiki) contains:
  - Research statements from all participants
  - Pointers to existing projects and ongoing work
  - Workshop presentations and breakout reports
  - Final presentation

- Workshop report
  - Will be published both on the wiki and a select journal

- Presentation of workshop results at a variety of meetings
  - Workshop on Workflows in Support of Large-Scale Science at HPDC06, GGF, AAAI06, Other forums
State of the Art: Useful Pointers

- Compilation of 1-pagers from this workshop (at workshop wiki site)
- SIGMOD record special issue Special Section on Scientific Workflows, September 2005 (see editorial + survey papers)
  - http://www.sigmod.org/record/issues/0509
- Journal of grid computing special issue
- [Yu & Buyya 05] survey
- Upcoming book on “Workflows in e-Science” by Gannon, Deelman, Taylor, Shields
- Upcoming book on “Workflows” Ian Hunter
- Web sites with pointers to workflow systems
  - http://www.extreme.indiana.edu/swf-survey/
- GGF workflow research activity and its web site
  - http://www.isi.edu/~deelman/wfm-rg/
- ICD’06 + HPCD’06 workshops on workflows
Four Proposed Topics for Workshop Discussions

- **Applications and requirements**
  - What are the requirements of future applications? What new capabilities are needed to support emerging applications?

- **Dynamic workflows and user steering**
  - What are the challenges in supporting dynamic workflows that need to evolve over time as execution data become available? What kinds of techniques can support incremental and dynamic workflow evolution due to user steering?

- **System-level management**
  - What are the challenges in supporting large-scale workflows in a scalable and robust way? What changes are needed in existing software infrastructure? What new research needs to be done to develop better workflow management systems?

- **Data and workflow descriptions**
  - How can workflow descriptions be improved to support usability and scalability? How to describe data produced as part of the workflows? What provenance information needs to be tracked to support scalable data and workflow discovery?
Four Discussion Topics

- **Applications and requirements**
  - What are the requirements of future applications? What new capabilities are needed to support emerging applications?

- **Dynamic workflows and user steering**
  - What are the challenges in supporting dynamic workflows that need to evolve over time as execution data become available? What kinds of techniques can support incremental and dynamic workflow evolution due to user steering?

- **Data and workflow descriptions**
  - How can workflow descriptions be improved to support usability and scalability? How to describe data produced as part of the workflows? What provenance information needs to be tracked to support scalable data and workflow discovery?

- **System-level management**
  - What are the challenges in supporting large-scale workflows in a scalable and robust way? What changes are needed in existing software infrastructure? What new research needs to be done to develop better workflow management systems?
The Application Drivers

- Exponential growth in Compute(18), Sensors(18?), Data storage(12), Network(8) (doubling time in months); performance variable in practice (last mile for networks)
- Growth of Science not same exponential
  - Perceived importance of workflows in accelerating pace of scientific advances
- Ground discussions in scenarios or in application descriptions that lead to CS requirements
  - Spans all NSF directorates
  - Astronomy (multi-wavelength VO), Biology (Genomics/Proteomics), Chemistry (Drug Discovery), Environmental Science (multi-sensor monitors as in NEON), Engineering (NEES, multi-disciplinary design), Geoscience (Ocean/Weather/Earth(quake) data assimilation), Medicine (multi-modal/instrument imaging), Physics (LHC, Material design), Social science (Critical Infrastructure simulations for DHS) etc.
Complex scientific methodology produces modern scientific results
Application Requirements I

- **Reproducibility** core to scientific method and requires rich provenance, interoperable persistent repositories with linkage of open data and publication as well as distributed simulations, data analysis and new algorithms.
  - Distributed Science Methodology publishes all steps in a new electronic logbook capturing scientific process (data analysis) as a rich cloud of resources including emails, PPT, Wikis as well as databases, compiler options, build time/runtime configuration…
    - Need to separate wheat from chaff in implicit electronic record (logbook) keeping only that required to make process reproducible; need to be able to electronically reference steps in process;
    - Traditional workflow including BPEL/Kepler/Pegasus/Taverna only describes a part of this
    - Abstract model of logbook becomes a high level executable meta-workflow

- Multiple collaborative heterogeneous interdisciplinary approaches to all aspects of the distributed science methodology inevitable; need research on integration of this diversity
  - Need to maximize innovation (diversity) preserving reproducibility
Application Requirements II

- Interdisciplinary science requires that we federate ontologies and metadata standards coping with their inevitable inconsistencies and even absence
- Support for curation, data validation and “scrubbing” in algorithms and provenance;
  - QoS; reputation and trust systems for data providers
- Multiple “ibilities” (security, reliability, usability, scalability)
- As we scale size and richness of data and algorithms, need a scalable methodology that hides complexity (compatible with number of scientists increasing slowly); must be simple and validatable
- Automate efficient provisioning, deployment and provenance generation of complex simulations and data analysis; support deployment and interoperable specification of user’s abstract workflow; support interactive user
- Support automated and innovative individual contributions to core “black boxes” (produced by “marine corps” for “common case”) and for general user’s actions such as choice and annotation
Four Focus Areas

■ **Applications and requirements**
  - What are the requirements of future applications? What new capabilities are needed to support emerging applications?

■ **Dynamic workflows and user steering**
  - What are the challenges in supporting dynamic workflows that need to evolve over time as execution data become available? What kinds of techniques can support incremental and dynamic workflow evolution due to user steering?

■ **Data and workflow descriptions**
  - How can workflow descriptions be improved to support usability and scalability? How to describe data produced as part of the workflows? What provenance information needs to be tracked to support scalable data and workflow discovery?

■ **System-level management**
  - What are the challenges in supporting large-scale workflows in a scalable and robust way? What changes are needed in existing software infrastructure? What new research needs to be done to develop better workflow management systems?
Vision

- Enable and accelerate distributed scientific methodology via rapid REUSE and EXPLORATION.
  - The Process-driven Scientific Method has exploration and evolution at its heart. Change is the norm.
  - The Processes are artefacts that change and assets to leverage. Workflows are a critical form of process
  - A technical support environment for the SM has to create a stable platform for science YET support the change in Science.
  - Get over it.
• Learn and represent the utility and effectiveness of algorithms
• Extract information from discovery processes (i.e. action history)
• Enable domain scientists to perform data exploration
• Workflow adaptation and evolution
• Recommend appropriate modifications to the user
• Validation of algorithms in dynamic environments
• Intelligent planning, scheduling, optimization and repair of workflows
• Researcher definable trust and security in distributed environments

• Support for effective user creation and adaptation of workflows
  - Semantic descriptions
  - Knowledge based guidance
• Support for heterogeneous tools
• Effective user interfaces that balance flexibility and ease of use

• Suitable models for the automated representation of action history and data provenance
  - Reproducability
  - Share Results
• User friendly explanation of workflow enactment
Integrative Science

- Understand the exploratory process
  - Models of collaboration
  - Models of change
  - Models of events
  - Models of workflow execution
  - Models of users
  - Models of actions on change responding to events
Resource Focus

- Where should the focus be?

- Development of workflow knowledge resources

- Recognize that workflows are sharable commodities

- Requires an interoperable infrastructure that treats workflows as “first class citizens”

- Support real collaboration between CS and domain scientists
Four Focus Areas

- **Applications and requirements**
  - What are the requirements of future applications? What new capabilities are needed to support emerging applications?

- **Dynamic workflows and user steering**
  - What are the challenges in supporting dynamic workflows that need to evolve over time as execution data become available? What kinds of techniques can support incremental and dynamic workflow evolution due to user steering?

- **Data and workflow descriptions**
  - How can workflow descriptions be improved to support usability and scalability? How to describe data produced as part of the workflows? What provenance information needs to be tracked to support scalable data and workflow discovery?

- **System-level management**
  - What are the challenges in supporting large-scale workflows in a scalable and robust way? What changes are needed in existing software infrastructure? What new research needs to be done to develop better workflow management systems?
Vision Statement

- Place a focus on Science and Engineering Research Processes / The Overall Scientific Lifecycle
  - Faithful capture of provenance
  - Analysis/abstraction from provenance to capture understanding and to instantiate new processes
  - Data, Model, Process, and Provenance are all first class products / content

- Benefits
  - Process is as sharable as data
  - One can navigate between the literature and the low-level executable processes and back, between scientific description and engineering flows
  - The gap between idea and execution is gone, the gap between provenance and instantiating new processes is gone
  - Workflow models/ provenance are a useful simplifying language for understanding large scale, flexible systems
  - Process descriptions can be used for building workflow-based ‘applications’
Research Challenges

- End-to-end process descriptions/provenance
  - Mechanisms for interacting with/understanding/using provenance
  - The study of scientific processes
- Robust engine(s) optimized for different purposes (desktop, streaming, event-driven, grid (large data, large computation), interactive, dynamic, workflow-by-example, ...)
  - Migrating processes across ilities
- Discovery of common workflow patterns and primitives
- Mine/query/link/reuse processes and process patterns and primitives
- Multi-layer workflows and automated/scaffolded workflow mapping
  - From scientific to engineering descriptions and vice versa
  - The ability to move workflows from interactive to batch, local to remote, etc.
- Workflow and provenance as content (versionable, sharable, organizable, discoverable, securable)
- Workflow in VO context (science model, policies (privacy, security, resource use, ...), ...)

NSF Workshop on Challenges of Scientific Workflows, May 1-2 2006
Four Focus Areas

■ Applications and requirements
  • What are the requirements of future applications? What new capabilities are needed to support emerging applications?

■ Dynamic workflows and user steering
  • What are the challenges in supporting dynamic workflows that need to evolve over time as execution data become available? What kinds of techniques can support incremental and dynamic workflow evolution due to user steering?

■ Data and workflow descriptions
  • How can workflow descriptions be improved to support usability and scalability? How to describe data produced as part of the workflows? What provenance information needs to be tracked to support scalable data and workflow discovery?

■ System-level management
  • What are the challenges in supporting large-scale workflows in a scalable and robust way? What changes are needed in existing software infrastructure? What new research needs to be done to develop better workflow management systems?
Infrastructure Needed

- We need a workflow infrastructure in order to transform both application science and computer science
  - Would share the cost of shared capabilities (e.g. provenance)
  - Need a community process to define the details

- Infrastructure components
  - A common workflow back-end
  - Workflow language(s) and tool(s) built on this
  - Infrastructure will evolve as research progresses
Research Areas

Many issues are unresolved today
  • Provenance throughout all layers
  • Data-flow aware services provided by the middleware
  • Building interoperable and composable interfaces
  • Automatic provisioning of software and resources
  • Scalability, heterogeneity, reliability, …
Summary of Workshop Discussions

- Workflows are “recipes” for Cyberinfrastructure computations
- Reproducibility, an important requirement in scientific method
  - Scientific and engineering reproducibility
  - Increasingly challenging to capture
- Science has exploratory and evolutionary nature
  - Need to support this dynamic nature
  - Workflows are key enabling technologies
  - Dynamic evolving user-steered workflows are the norm in science
- Provenance is key to reproducibility
  - Provenance and process description are an end-to-end process
  - At all levels of abstraction -- science domain level and system level
- Stable and common software platforms
  - Infrastructure will evolve, stability is crucial for reproducibility
Summary of Workshop Recommendations

- Embed CS experts in the application domains
- Support for longer term, stable (5+ year) collaborations + projects
  - Based on experience in current CS/Science collaboratories
- Capture best practices – provide coordination framework
- Need explicit representations for workflows (at different levels of abstraction)
- Need a framework where various workflow tools can interoperate
  - More work needs to be done on Workflow System: construction, planning, execution, debugging
- Incorporate HCI research: how to present execution, options, recommendations and trails of what happened
- Incorporate social sciences, organizational sciences and business workflow communities
  - Analyze and capture of scientific process
- Foster follow-up cross-cutting workshops and meetings
- Need to invest in this important area for the future of science: Just as scientists would face a crisis if the data they collect is lost, they will face a great crisis if the analyses are lost because their computational workflows are not captured