Outline

Background

Design

Workflow Systems

AI Workflows

Future
Tutorial Schedule

9:00    Part I: Background
  • General background on computational workflows
9:30    Part II: Designing Workflows
  • Casting complex applications as workflows
10:00   Coffee Break
10:20   Part III: Creating Workflows in practice
  • Specifying high-level workflows using Wings
11:00   Part IV: Executing Workflows in practice
  • Automatic mapping and execution of workflows with Pegasus
11:20   Demonstration
11:40   Part V: AI Workflows
  • Examples of AI workflows including machine learning and natural language processing
12:10   Part VI: A survey of scientific workflow systems
  • Overview of other research on scientific workflows
12:30   Part VII: The Future
  • Ongoing work and open challenges relevant to AI research
Reading About Workflows

AAAII-08 Tutorial on
Computational Workflows for
Large-Scale Artificial Intelligence Research

Part I:
Background
Scientific Collaborations: Publications

[Barabassi 2005]
Computing and the Future of Science
Science is Undergoing a Significant Paradigm Change

- Entire communities are collaborating and pursuing joint goals
  - Astronomy (SDSS, NVO), Biology (BIRN), Environmental Science (NEON, OOI), Engineering (NEES), Geoscience (SCEC, GEON), Medicine (CaBIG), Physics (LHC, LIGO), etc.

- Instruments, hardware, software, and other resources shared (TeraGrid, OSG, NMI)

- Data shared and processed at large scales

- Shared distributed collaborations: “Collaboratories”
Sharing Data Collection: LIGO  (ligo.caltech.edu)
Sharing Computing Resources

TeraGrid: A National Production CI Facility

Phase I: 2001-2004 Design, Deploy, Expand
Phase II: 2005-2010 Operation & Enhancement

20+ Distinct Computing Resources
(>100TF today; >200TF by 2007)
100+ Data Collections

TeraGrid Users

Blue: 10 or more PI’s
Red: 5-9 PI’s
Yellow: 2-4 PI’s
Green: 1 PI

1000 projects, 4000 users
Integrating Diverse Models of Complex Scientific Phenomena

Seismicity

Paleoseismology

Local site effects

Geologic structure

Faults

Stress transfer

Crustal motion

Crustal deformation

Seismic velocity structure

Rupture dynamics

Seismic Hazard Model
Scale in AI

Multi-disciplinary experiments  Large-scale models  Shared, large-scale resources

- While many sciences benefit from large-scale processing…
- … AI research is largely done in small scale
  - Typically confined to desktop computations with modest data sizes

Model integration leads to new discoveries
“Cyberinfrastructure” Sharing in Scientific Collaboratories

- Distributed environment with selective sharing
  - people, data, computing, code, instruments

- Complex analysis processes
  - Need to combine individual algorithms into valid end-to-end integrated analysis

- Large resource requirements
  - computing and data

- Shareable, reproducible results and analysis process
  - Need to keep track of how analysis was generated

- Evolving requirements and models
  - Scientific knowledge and resources are always changing

- Very dynamic environment
  - Models (code), availability of computing resources, data, etc
Common Cyberinfrastructure Layers

- Portals
- Data Services
- Resource Access
- Resource Sharing
- Application Tools

USC Information Sciences Institute
Yolanda Gil (gil@isi.edu)
AAAI-08 Tutorial
What Cyberinfrastructure is Missing

- Current CyberInfrastructure is an enabler of a significant paradigm change in science
  - Distributed interdisciplinary data rich computational experimentation is leading to a transformative approach
- However:
  - Reproducibility, key to scientific practice, is threatened
    - Process (method/protocol) is increasingly complex and highly distributed
  - Exponential growth in Compute, Sensors, Data storage, Network BUT growth of science is not same exponential
    - Perceived importance of capturing and sharing process in accelerating pace of scientific advances
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

Report available at http://www.isi.edu/nsf-workflows06
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.

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

- Workflows provide a formalization of the scientific analysis
  - 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
  - Method is captured and can be reused by others at zero-cost
  - Guarantee of data “pedigree”

- Workflows are structures useful to manage computation
  - Workflow system can provide assistance, automation, records

- Objects of scientific discourse: collaboratively designed, assembled, validated, analyzed, evolved
Workflow Systems as Key Cyberinfrastructure Layer

Data Services

Workflow Systems

Resource Sharing

Application Tools

Resource Access

Portals

Portals

Portals
How Scientists Develop Complex Applications Today

- Scientists have high level requirements naturally stated in terms of the application domain
  - Ex: Obtain frequency spectrum for signal S in instrument I and timeframe T
- These requirements can be achieved by combining models
- Models are often complex in terms of size and HPC requirements
- So, scientists must be well trained on high performance/distributed computing (grids)
- First, they have to turn these requirements into combinations of executable jobs specified in detailed scripts
  - They must figure out which code generates desired products, which files contain it, physical location of the files, hosts that support execution given code requirements, availability of hosts, access policies, etc.
  - They have to be able to query grid middleware: metadata catalog, replica locator, resource descriptor and monitoring, etc.
- They must also oversee execution
  - Diagnose failures (code, memory, network, resource, etc) and design recovery strategies (replace resource, rearrange data, replace code, etc)
Workflow Management through Scripts

- Scripts that specify the control structure of the workflow to be executed
  - Generate input values to all application codes in the workflow from a starting input file
  - Determine the selection of application codes based on starting input file
  - Keep track of where new results come from (provenance)

- Scripts provide a common framework to compose models

- Scripts-based approaches are a first step in managing computation, used by many

- But…
Problems with Script-Based Approaches

- Adding a new requirement affects a lot of scripts
- Adding a new model (or a new version of a model) requires changes to starting input file and going through scripts by hand
  - Error prone process
- Ad-hoc data and execution management
  - Manually check whether intermediate data already exists
  - Metadata generated by scripts and passed around
  - To run workflow at other hosts, the scripts have to be changed to have the right file paths
- Customized interfaces created for non-experts to ensure the workflow is run correctly
Scientific Workflows

- Emerging paradigm for large-scale and large-scope scientific inquiry
  - Large-scope science integrates diverse models, phenomena, disciplines
  - “in-silico experimentation”

- Workflows provide a formalization of the scientific analysis
  - 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

- Workflow are structures useful to manage computation
- Collaboratively designed, assembled, validated, analyzed