Astrophysics Workflows in the Kepler System

Ilkay Altintas
Director, Scientific Workflow Automation Technologies Laboratory
San Diego Supercomputer Center, UCSD
The Big Picture: Supporting the Scientist

John Blondin, NC State Astrophysics
Terascale Supernova Initiative
SciDAC, DOE

From “Napkin Drawings” …

… to Executable Workflows

Conceptual SWF

Executable SWF
Scientific Workflow Systems

θ Combination of
  ♣ data management, integration, analysis, and visualization steps
  ♣ larger, automated "scientific process"

θ Mission of scientific workflow systems
  ♣ Promote “scientific discovery” by providing tools and methods to generate scientific workflows
  ♣ Provide an extensible and customizable graphical user interface for scientists from different scientific domains
  ♣ Support workflow design, execution, sharing, reuse and provenance
  ♣ Design frameworks which define efficient ways to connect to the existing data and integrate heterogeneous data from multiple resources

θ Make technology useful through user’s computer!!!
Kepler is a Scientific Workflow System

- ... and a cross-project collaboration
- 3rd Beta release (Jan 8th, 2007)

- Builds upon the open-source Ptolemy II framework

Ptolemy II: A laboratory for investigating design
KEPLER: A problem-solving environment for Scientific Workflow

KEPLER = “Ptolemy II + X” for Scientific Workflows
Kepler is a Team Effort

Other contributors:
- Chesire (UK Text Mining Center)
- DART (Great Barrier Reef, Australia)
- National Digital Archives + UCSD-TV (US)
- ...

Contributor names and funding info are at the Kepler website!!
Actors are the Processing Components

Actor
- Encapsulation of parameterized actions
- Interface defined by ports and parameters

Port
- Communication between input and output data
- Without call-return semantics

Model of computation
- Communication semantics among ports
- Flow of control
- Implementation is a framework

Examples
- Simulink (The MathWorks)
- LabVIEW (from National Instruments)
- Easy 5x (from Boeing)
- ROOM (Real-time object-oriented modeling)
- ADL (Wright)
- ...

Actor-Oriented Design
Some actors in place for…

- Actors that wrap **CCA** components
- Generic **Web Service** Client and Web Service Harvester
- Customizable **RDBMS** query and update
- **Command Line** wrapper tools
  - local, ssh, scp, ftp, etc.
- Some **Grid** actors
  - Globus Job Runner, GridFTP, Certificate Generator
- **SRB** support
- Native **R** and **Matlab** support
- Interaction with **parameter sweep** tools
  - Nimrod and APST
- Communication with **ORBs** through actors and services
- **Image Processing**, **Gridding**, **Visualisation** Support
- Textual and Graphical **Output**
- Even wrapping subsystems, such as **SCIRun**
- …more generic and domain-oriented actors…
Vergil is the GUI for Kepler

- Actor ontology and semantic search for actors
- Search -> Drag and drop -> Link via ports
- Metadata-based search for datasets
Kepler can be used as a batch execution engine

- **Configuration phase**
- **Subset**: DB2 query on DataStar

- **Interpolate**: Grass RST, Grass IDW, GMT…
- **Visualize**: Global Mapper, FlederMaus, ArcIMS
TSI Workflow For Doug Swesty

Acknowledgements:
Terence Critchlow, Xiaowen Xin and the rest of the SDM SPA team
Goals: integrate and deploy software-based solutions to efficiently and effectively manage large volumes of data generated by scientific applications

SDM Center focuses on the application of known and emerging data management technologies to scientific applications.

http://sdm.lbl.gov/sdmcenter/

Funded under SciDAC-1
TSI Workflow-2 (D. Swesty)
Submit batch request at NERSC

Check job status

Identify new complete files

Transfer files to HPSS

Transfer completed correctly

Transfer files to SB

Transfer completed correctly

Delete file

Update web page

Extract
- Get Variables
- Create Chem vars
- Create neutrino vars
- Remap coordinates
- Derive other vars
- Write diagnostic file

Generate plots
- Tool-1
- Tool-2
- Tool-3
- Tool-4

Generate movie

Generate thumbnails
TSI-2 Executable Workflow Screenshot
- **Actors linked together with relations**
- **Relations link ports on actors**
- **Tokens contain information passed between actors**
- **Composite actors provide a layer of abstraction**
TSI-2 Web Interface for Monitoring
TSI-2 Workflow Running Interface
Parallelism and Generality

- Every actor runs in its own thread
- Actors start running as soon as it has enough data
- File transfers to HPSS can take place simultaneous with transfers from HPSS

This workflow uses a small number of actors repeatedly to deliver a complex behavior

- 160 instances an actors
- 18 different types of actors
- ~100 Expression actor instances
- 13 boolean switch actor instances
- 13 array manipulation actor instances
- 8 ssh actor instances
- < 30 instances of other actors (ex: sleep, file I/O, etc)
Our collaborative efforts with the SDM Center over the last few years, has enabled us to publish several peer reviewed papers, and abstracts that were an important factor in maintaining current funding in the low-dose radiation research program (http://lowdose.tricity.wsu.edu/). … Having access to SDM Center scientists and their automated workflow and parallel processing tools over the next two years will be critical for the identification and characterization of regulatory element profiles of IR-responsive genes and will provide valuable understanding of the genetic mechanisms of IR-response and should provide powerful biological indicators of genetic susceptibilities for tissue and genetic damage.

“We are finally seeing some nice payoffs in terms of easy-to-use computational chemistry software with some unique capabilities. The framework illustrates nicely the interplay between technology and applications - including the compute software, the middleware, and the grid computing capabilities.”

During SciDAC I, members of the Terascale Supernova Initiative (TSI), now being expanded to PSI, collaborated extensively with members of your team. And the benefits were palpable. The successful deployment of a scientific workflow management and automation tool, which arose out of a fruitful collaboration between Doug Swesty and Eric Myra of Stony Brook and Terence Critchlow of LLNL, is one example. … Moreover, others of your effort (e.g., Mladen Vouk) are based at PSI partner sites and engaged in helping some of our application scientists (in this case, John Blondin), which will further enhance our overall research exchange.

The CIPRES project has as a key goal the creation of software infrastructure that allows developers in the community to easily contribute new software tools, … The modular nature of Kepler met our requirements, as it is a JAVA platform that allows users to construct linear, looping, and complex workflows from just the kinds of components. The CIPRES community is developing. By adopting this tool, we were able to focus on developing appropriate framework and registry tools for our community, and use the friendly Kepler user application interface as an entrée to our services. We are very excited about the progress we have made, and think the tool will be revolutionary for our user base.

It can help you too!

- Matthew A. Coleman, Bioscience Program, Lawrence Livermore National Laboratory, 2005
- Kim K. Baldridge, PI, Resurgence project, 2004
- Anthony Mezzacappa, PI, DOE SciDAC Petascale Supernova Initiative, 2005
- Mark A. Miller, PI, NSF CIPRES project, 2006
To Sum Up

... is funded under SciDAC-1 and 2

- has a scientific process automation thrust area called SPA
- SPA is a collaborator in the Kepler consortium
- SPA has developed two TSI workflows

... is an open-source system and collaboration

- is a 2.5 year-old project
- grows by application pull from contributors
- released Beta3.0 on Jan 8th, 2007

There is ongoing work on ocean and ecological observatories…

- See http://reap.ecoinformatics.org
Questions...

Thanks!

Ilkay Altintas  
altintas@sdsc.edu 
+1 (858) 822-5453  
http://www.sdsc.edu