Medical-Image Processing Workflow Support on EGEE with Taverna

Ketan Maheshwari, Paolo Missier, Carole Goble, Johan Montagnat

UNS / CNRS-I3S Laboratory and The University of Manchester

August 3, 2009

Overview

- Optimize the enactment of Data Intensive Workflows on Grid Infrastructures.
- Extend Taverna user community to biomedical applications.
- Grids enabling data intensive medical image analysis applications.
- This work attempts to achieve this objective by employing a ‘Grid plugin’ to the Taverna workbench.
- This presentation details the Taverna gLite plugin Design, Implementation and its Usage.

Taverna

- A popular scientific workflow manager with 1000+ strong user base.
- Advanced enactment capabilities including pipelining and data parallel enactment mode.
- Extendible Architecture of Taverna enables custom plugin development.
- Sophisticated User Interface but lack of grid integration.
- Taverna workbench ease the access to Grid infrastructure.

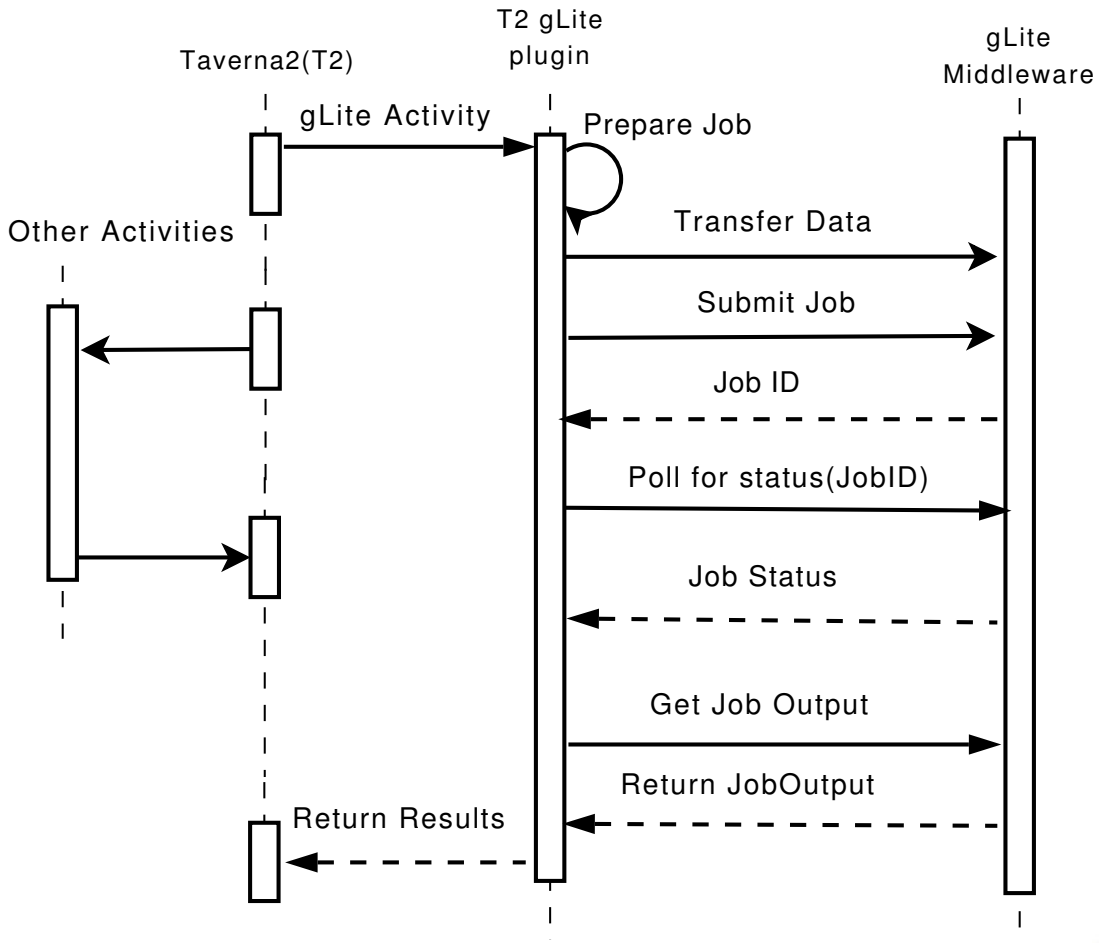
EGEE and gLite

- EGEE (Enabling Grids for E-sciencE) is a Premier European Grid Infrastructure.
- 125 Virtual Organizations, 9000+ users in 50 countries, $\tilde{2}$0 petabyte of storage and $\tilde{8}$0000 processor cores.
- Computation abstracted to Compute Elements (CE) and Storage to Storage Elements (SE).
- gLite is a de facto middleware program to access the EGEE batch system environment.

Design Challenges

- The main design challenge constitutes of coupling Taverna and EGEE environments.
- Transforming asynchronous Taverna calls to the batch-oriented, poll-based EGEE is a challenge.
- Overcome the Grid reliability issues.

Taverna EGEE Interaction via gLite Plugin



Implementation

- Process Description: Auto-generation of gLite Job Description Language.
- Data Transfer: Auto-generation of wrapper script.
- Job Status Polling : Configurable Frequency.

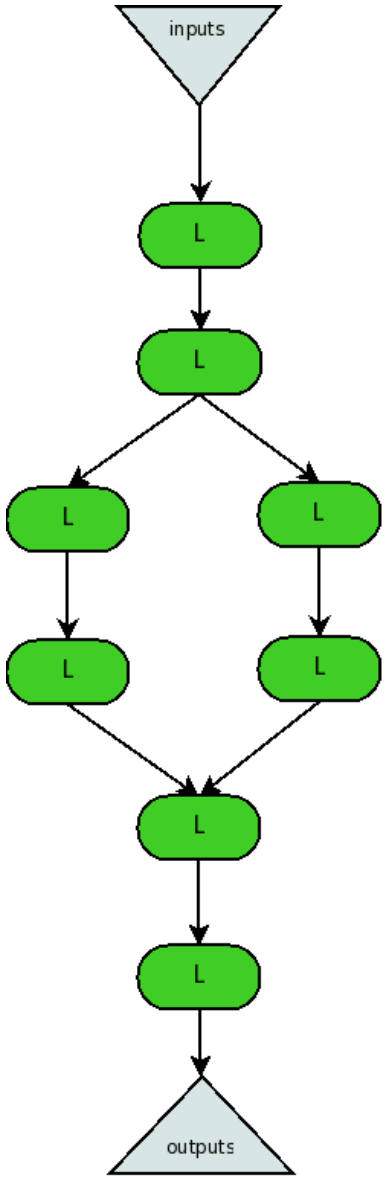
Addressing Reliability

- Job Resubmission on Error, long Wait or Abort state.
- Round Robin selection of Resource Brokers.
- Reliable data transfer: Repeat transfer in case of failure, rotate Storage Elements.

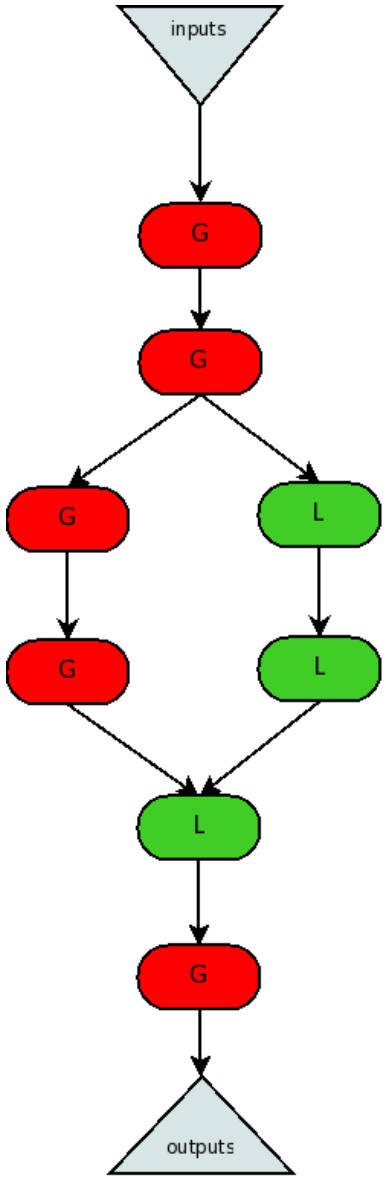
Usage

- Workflow Composition panel of Taverna provides a gLite processor.
- Configurable properties of the gLite processor.
- Automatic proxy delegation.
- Configurable polling frequency.
- Readily alterable execution mode.

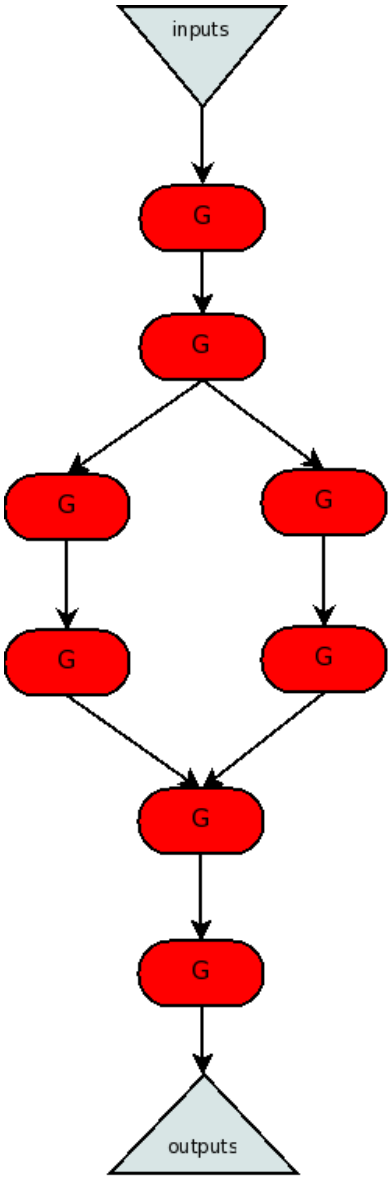
Execution Modes: Pure Local



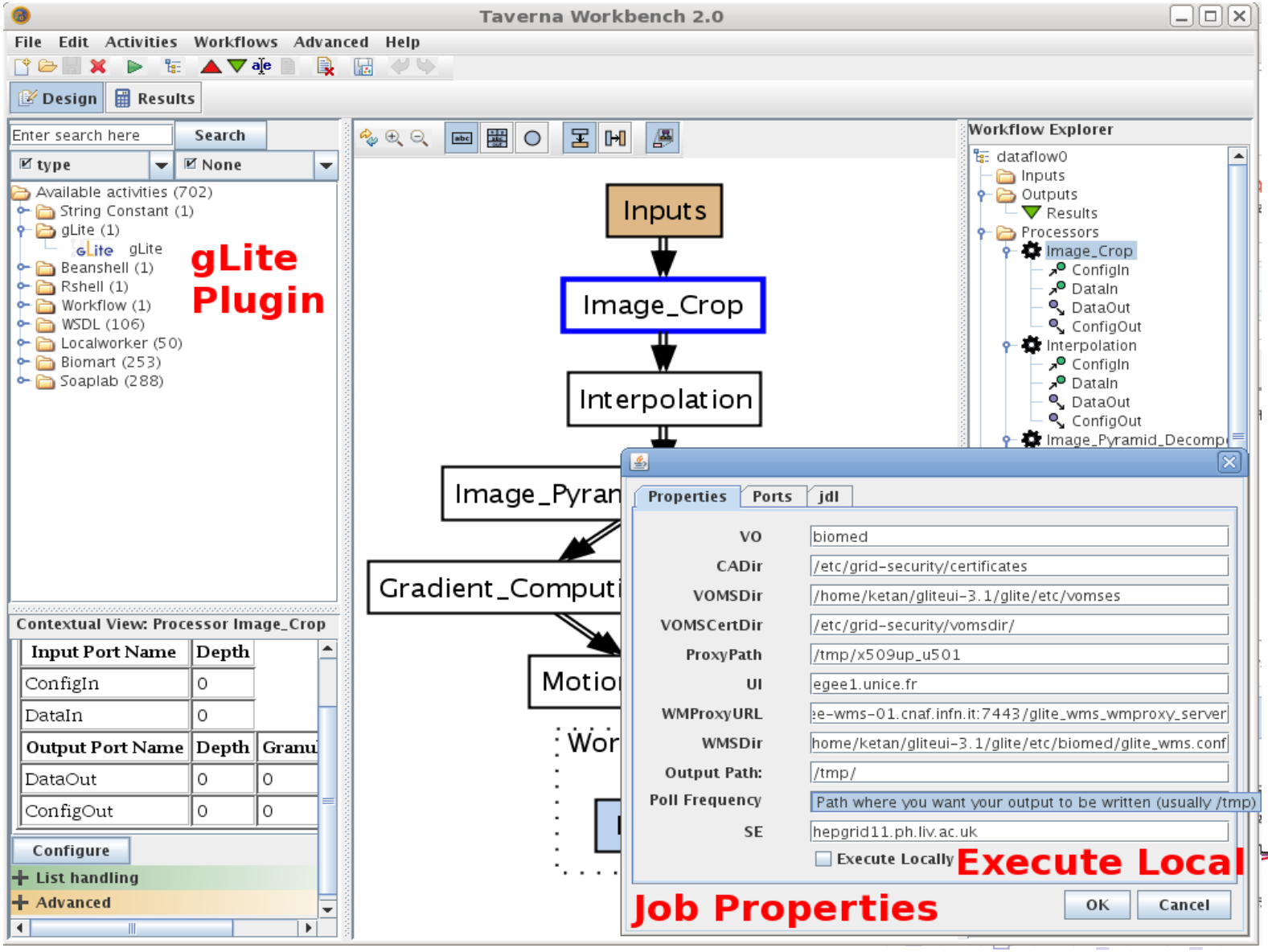
Execution Modes: Local/Grid Mixed



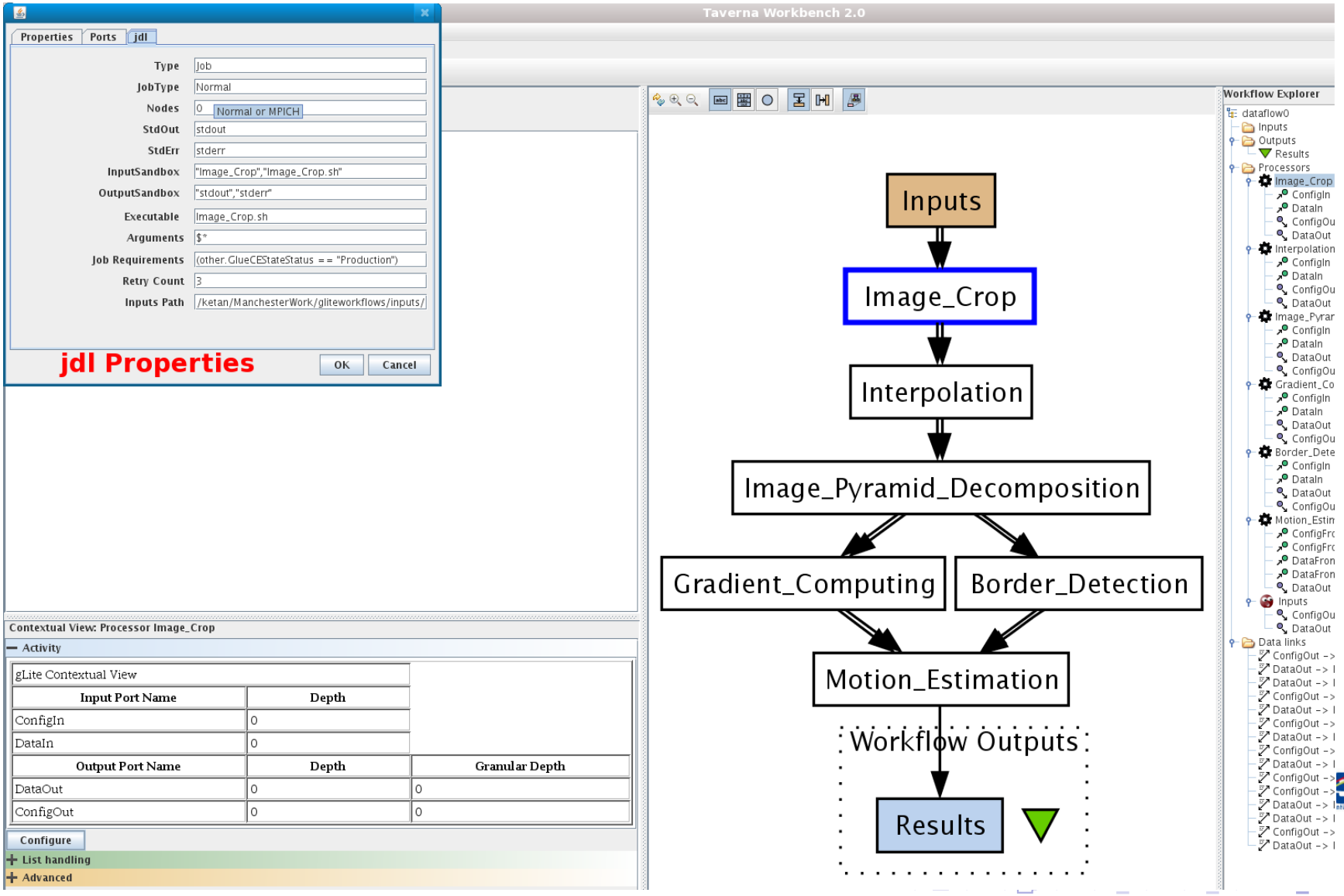
Execution Modes: Pure Grid



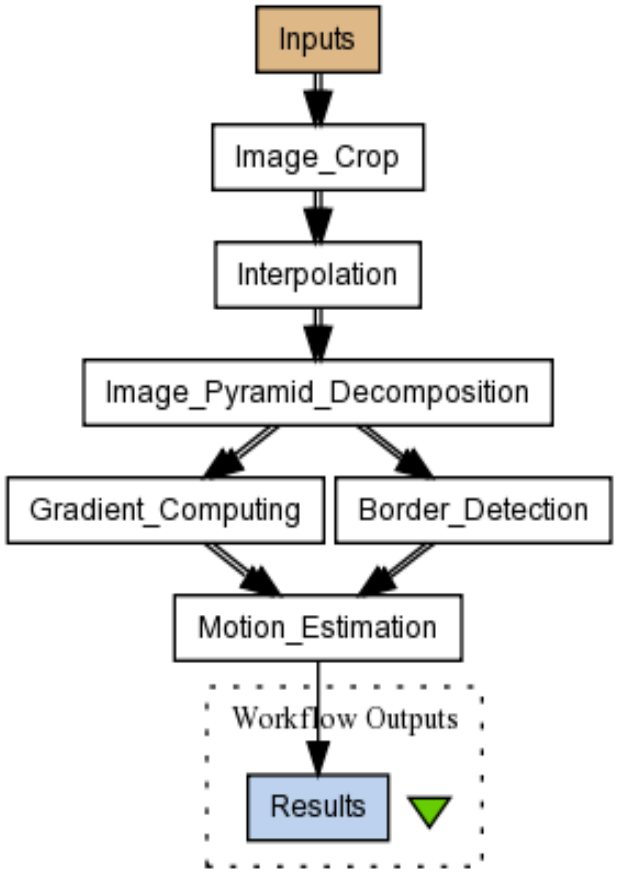
Screenshots(1)!



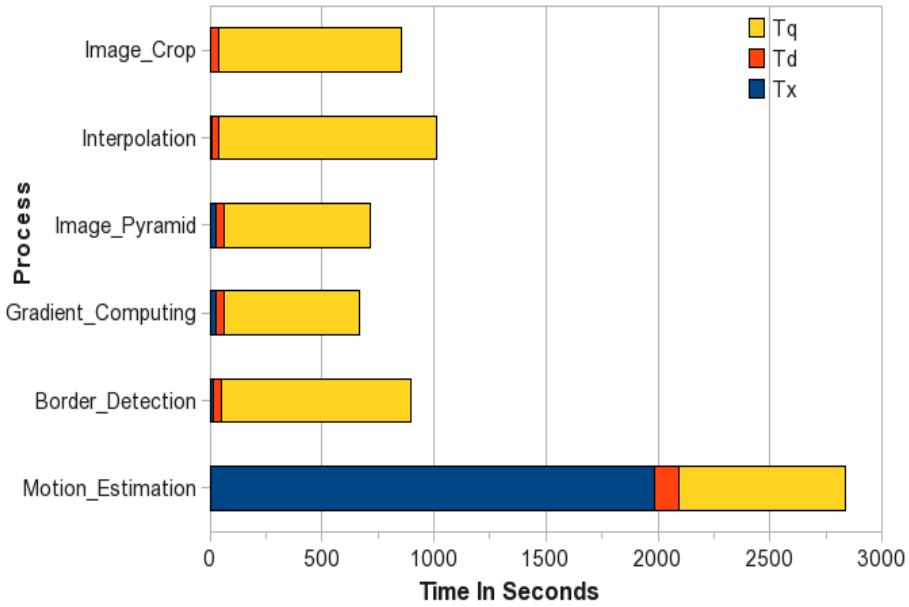
Screenshots(2)!



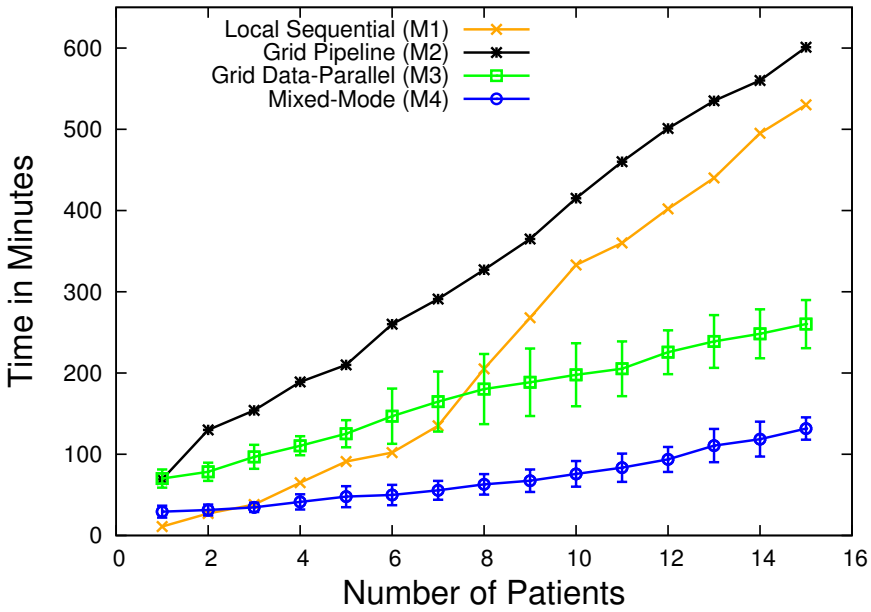
The Workflow and Results (1)



Orange(Tq)=queueing overhead,
Red(Td)=data transfer overhead,
Blue(Tx)=execution time



The Workflow and Results (2)



- Patients: 15
- Data: 20-30M/patient
- Peak Load: 65 concurrent threads

Conclusions

- gLite plugin is one of the first development in Taverna to interface with the grid and *the* first with EGEE.
- Emphasis on ease of workflow composition and Grid execution & data transfer reliability.
- 'Mixed-mode' enables easy empirical tests of data-intensive workflows.
- Can be easily applied to workflows from other domains involving data pipelines.

Thanks!, Questions?

This work is supported by the French ANR GWENDIA project under contract number ANR-06-MDCA-009. We are thankful to *my*Grid developers team for their excellent technical support and CREATIS (CNRS-INSERM) laboratory for providing application components. We thank Oleg Sukhoroslov for providing jLite java API.