Workflow Support on EGEE with Taverna

Ketan Maheshwari

UNS / CNRS-I3S Laboratory

July 28, 2009

Overview

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

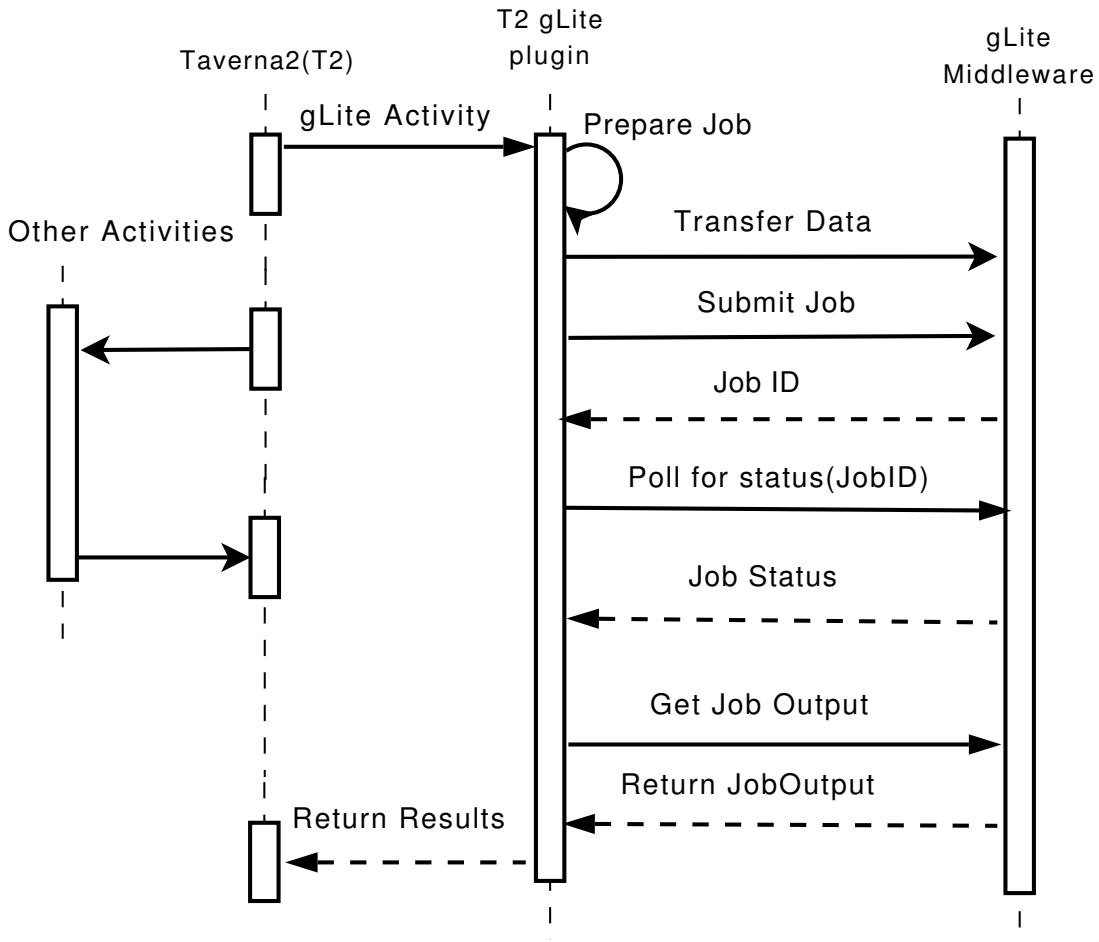
Taverna

- A popular scientific workflow manager with 1000+ strong userbase.
- 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.

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

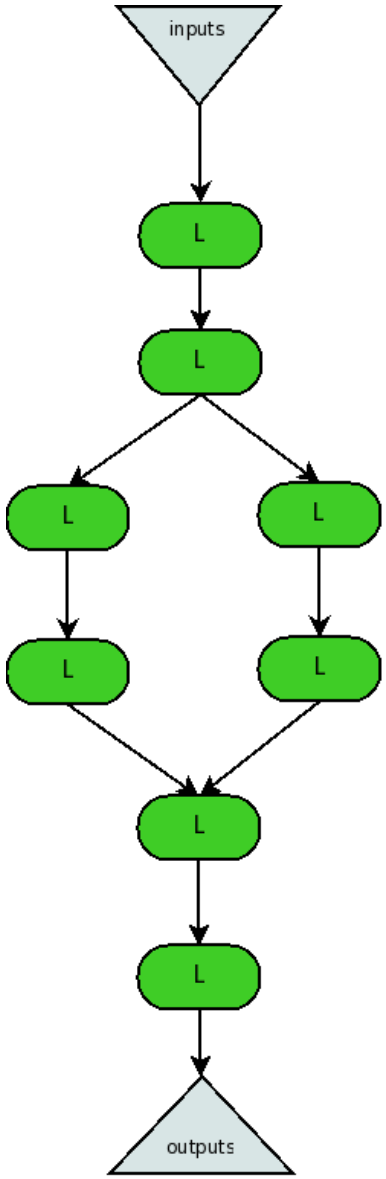
Addressing Reliability

- 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 Element.

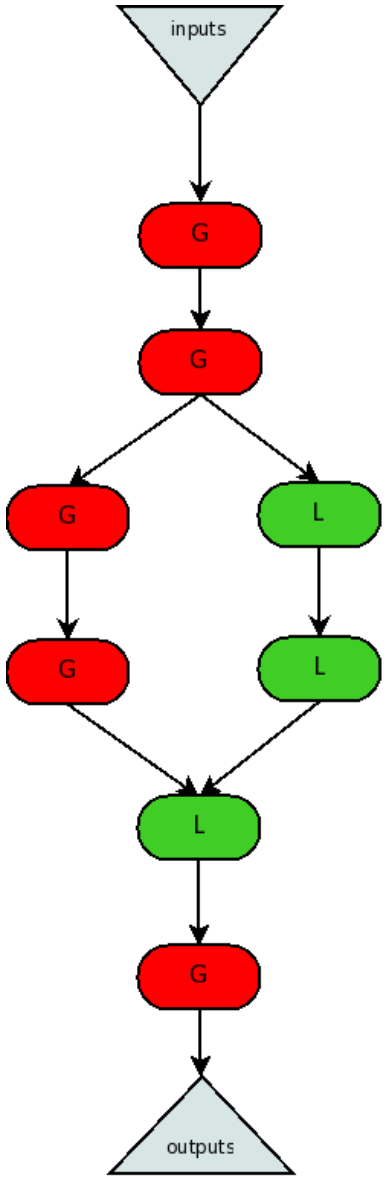
Usage

- 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.

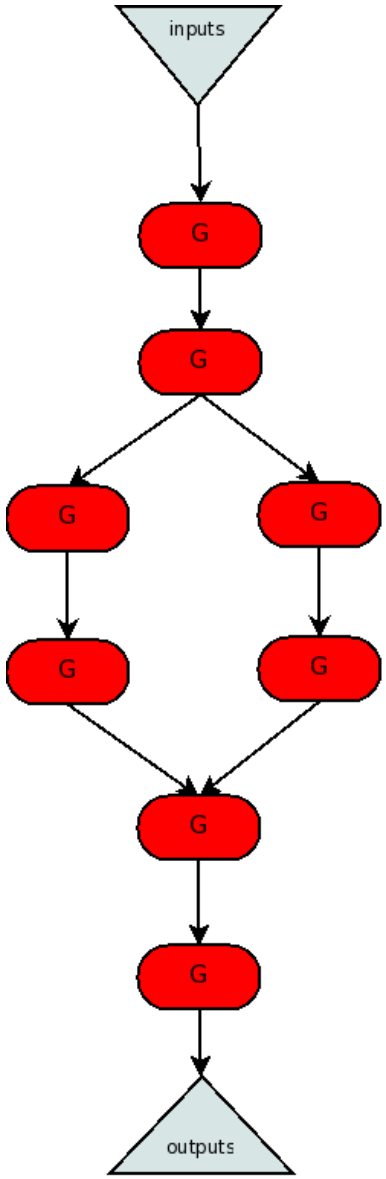
Execution Modes: Pure Local



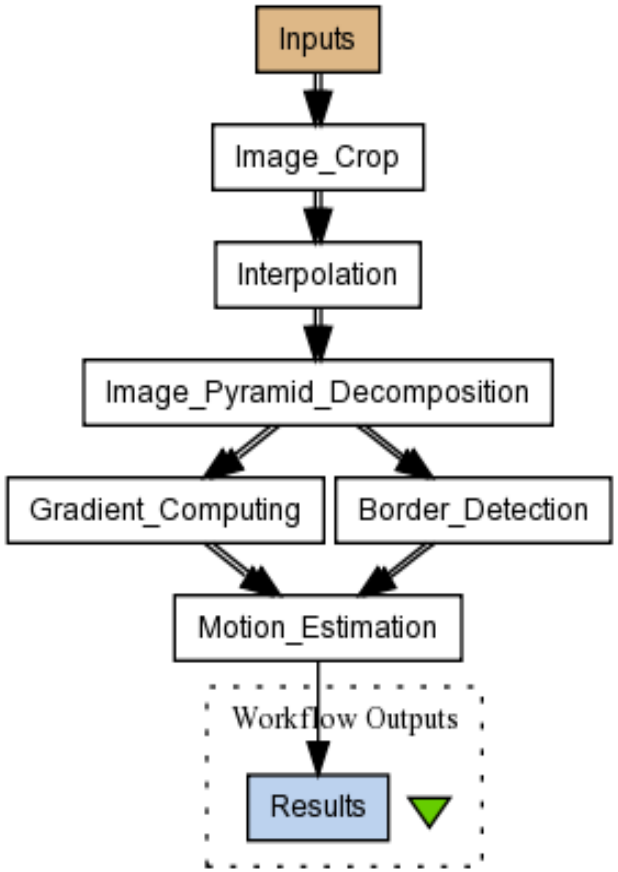
Execution Modes: Local/Grid Mixed



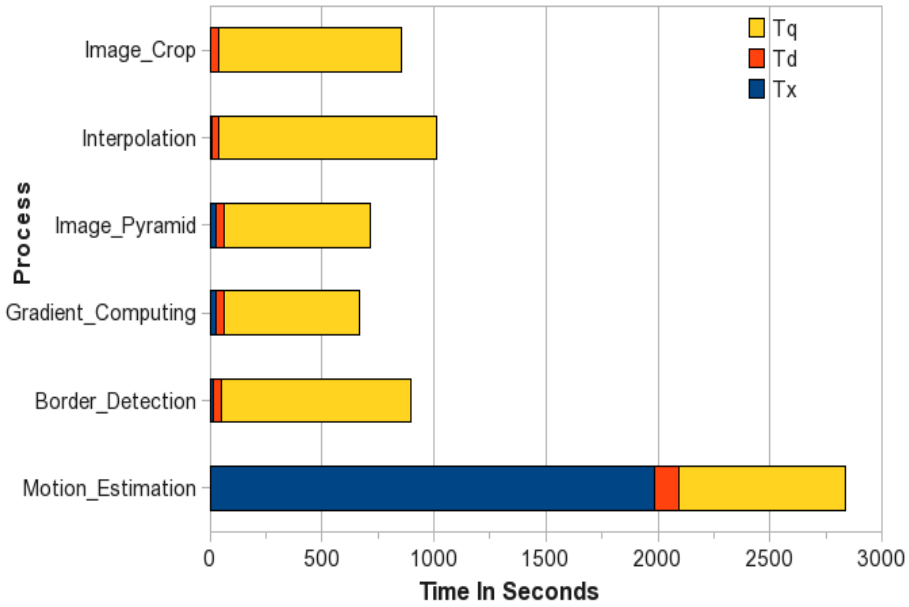
Execution Modes: Pure Grid



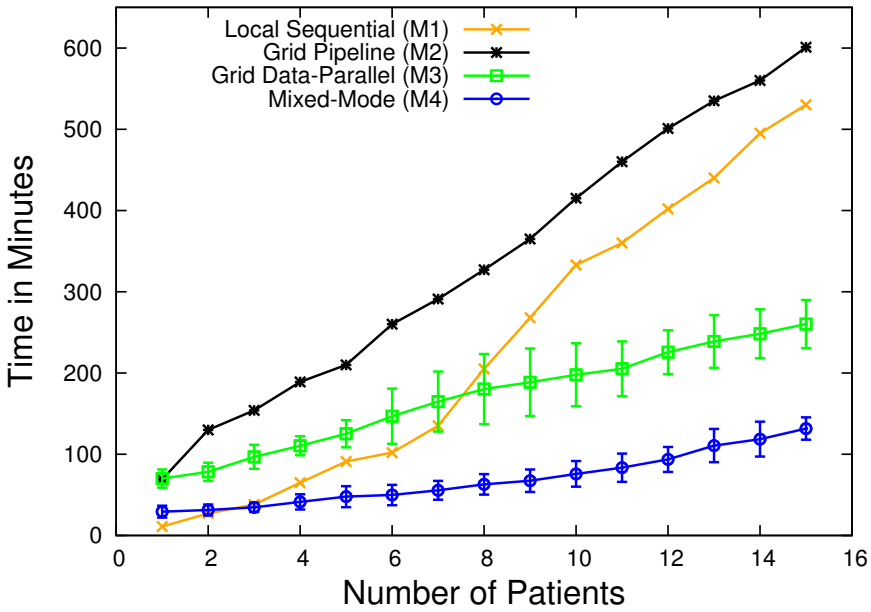
The Workflow and Results (1)



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.
- Can be easily applied to workflows from other domains involving data pipelines.

Thanks!

Questions!?!