EGI User Forum 2011

Contribution ID: **23** Type: **Poster**

Enabling High Throughput Computational Chemistry

Monday, 11 April 2011 09:00 (8 hours)

Overview

The project aim to provide a toolbox for ease the integration of a large scale computing infrastructure into computational chemistry high throughput pipelines.

The proposed poster shows how the high throughput computing framework 'gc3pie' has been used to implement automatic pipelines for generating online molecules database from GAMESS-US analysis.

The overall solution combines the analysis support that fully leverage the Swiss National Computing Infrastructure (SMSCG) based on ARC middleware, and the molecule database publishing.

GC3pie is a python-based overlay grid that can be used to integrate grid infrastructures and provide at application integration level a simple set of programming interfaces that can be used to develop scalable computing pipelines. It is an application centric framework that could be used to build e-science programming environments.

Plan for integrating GC3pie framework, as well as the scientific pipeline developed with it, into EGI is also presented

Impact

For several scientific usecases, the access to a very large computing infrastructure sounds beneficial but poses non-trivial problems on how to enable and control a scientific pipeline on such an infrastructure.

GC3pie provides an easy and programmable support for integrating computing infrastructures; it provides a controls over an high throughput executing with a minimal impact on the scientific pipeline.

It compartmentalize the access to the underlying computing infrastructure thus exposing through its high level programming interfaces only those minimal control required to express the executing of a given set of applications.

The main advantage is that through GC3pie is possible to develop pipelines focusing more on what characterize the pipeline (hopefully from the scientific point of view) rather than controlling the behaviour of the underlying computing infrastructure.

The programmable Task model allows to develop python-based pipelines (pretty much like workflows) and to control them thorough a series of event that could be triggered and handled during the pipeline lifetime.

Description of the work

The poster will present two main aspect of the current effort in enabling high throughput computational chemistry:

- GC3pie as a framework to easily integrate a large scale distributed infrastructure.
- The pipeline for generating online molecules databases from GAMESS-US analysis

The GC3pie framework has been mainly motivated by the need of programmatically integrate computing infrastructures into high throughput scientific pipelines.
The goal is to provide means for launching, controlling and post-process a very large number of jobs of various type and, at the same time, provide a programmatic abstraction for building scientific pipelines without having to deal directly with concepts like job, resource, batch system.
GC3pie provides an application centric programming model. It is a lightweight overlay grid that can be deployed on ARC-enabled client nodes. It also provides access to non grid-enabled resources (like a local or remote LRMS)
It represents the basic building block for e-science environment that could be built using the GC3pie abstractions.

The online molecule database generation system has been developed using the GC3pie libraries. It shows the added value of the programming model as well as the clear separation between the control of jobs execution (framed withing GC3pie) and the pipeline execution (which can be steered by the pipeline developer)

A pipeline can be compared to a workflow with the main difference that in GC3pie the pipeline is fully programmable and can be steered according to any event triggered during the pipeline lifetime.

GC3pie controls the access to the underlying computing infrastructure (the current implementation allows to integrate ARC connected resources) trying to optimize and encapsulate the access details.

SAGA is an example of a programmatic way of controlling computing grids; GC3pie provides similar functionality focusing more on the high level programming model (in fact both can be complementary with each other.

URL

http://code.google.com/p/gc3pie/

Conclusions

The GC3pie framework is an overlay grid that can be used to develop e-science environment fully integrated with a large scale computing infrastructure like EGI.

To fully enable Computational Chemistry on grids, one of the main challenge is to control high throughput executions.
Another is to develop scientific pipelines using known and flexible tools (like python).

Developing the online molecule database generation from GAMESS-US analysis has showed the added value of integrating both the access to the computing infrastructure as well as the control of high throughput pipelines from a programmatic perspective.

GC3pie can be considered as one of the building block for e-science environments.

Primary authors: Mr MURRI, Riccardo (UZH); Dr MAFFIOLETTI, Sergio (UZH)

Co-authors: Mr MONROE, Mark (UZH); Mr PACKARD, Mike (UZH)

Presenter: Dr MAFFIOLETTI, Sergio (UZH)

Session Classification: Posters

Track Classification: Poster