Contribution ID: **33** Type: **not specified**

GC3Pie: A Python framework for high-throughput computing

Thursday, 29 March 2012 14:45 (45 minutes)

Description of the Work

We 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 like EGI
- Examples of enabled high throughput usecases from various scientific domains

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 (access to CREAM-CE has been tested but not yet enabled in production). 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.

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 (being it aggregated by a grid middleware or not) trying to optimize and encapsulate the access details.
SAGA is another example of a programmatic way to control computing grids; while SAGA provides rich primitives to control and direct the execution of grid jobs and handle related data, GC3pie provides higher level programming interfaces to ease building grid applications and high throughput pipelines. Both are complementary with each other as SAGA provides interfaces to the grid infrastructure while GC3pie provides interfaces for the application and pipelines development.

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.

A key aspect in enabling scientific communities is to provide flexible mechanisms to control high throughput executions. Another is to develop scientific pipelines using known and flexible tools (like python).

GC3pie can be considered as one of the building block for e-science environments. GC3pie allows the scientist to focus on science instead of worrying about how to manage their jobs on the various grids. It provides a programmatic abstraction for building scientific pipelines without having to deal directly with concepts like authentication, grid middleware, batch systems, and individual system quirks. It is a flexible system supporting EGI resources.

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 large number of executions on such an infrastructure. GC3pie provides an easy and programmable support for integrating computing infrastructures; it provides means to control large number of executions and programmable interfaces to integrate such control in scientific pipelines. A pipeline is analogous to a workflow; the difference being that a pipeline is fully programmable and can be steered according to any event triggered during its lifetime.

Several usecases from different user communities have been enabled and relay on GC3Pie to access the national NGI-CH as well as EGI.

At the moment GC3Pie is used in production on the NGI-CH infrastructure for the following usecases:
Life sciences: Selectome (phylogeny), gmhc_coev (evolutionary biology), ROSETTA (system biology)
Computational Chemistry: GAMESS, GFIT_ABC
Computer cryptography: RSA768
Economic VO: gpremium, george

URL

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

Overview (For the conference guide)

GC3Pie is a suite of Python classes (and command-line tools built upon them) to aid in submitting and controlling batch jobs to clusters and grid resources seamlessly. GC3Pie aims at providing the building blocks by which Python scripts that combine several applications in a dynamic workflow can be quickly developed. GC3Libs, the main component of the GC3Pie framework, provides services for submitting computational jobs to Grids and batch systems and controlling their execution, persisting job information, and retrieving the final output. GC3Libs takes an application-oriented approach to batch computing. A generic Application class provides the basic operations for controlling remote computations, but different Application subclasses can expose adapted interfaces, focusing on the most relevant aspects of the application being represented.

Primary author: Dr MAFFIOLETTI, Sergio (UZH/GC3)

Co-authors: MURRI, Riccardo (UZH/GC3); ALEKSIEV, Tyanko (UZH/GC3)

Presenter: Dr MAFFIOLETTI, Sergio (UZH/GC3)

Session Classification: HTC/HPC

Track Classification: Software services for users and communities