Abstract

Encouraged by the success of the first EGEE BioMed data challenge battling malaria (WISDOM, http://wisdom.eu-egee.fr), the second data challenge was kicked off in April 2008 aiming at avian flu. The goal of the second data challenge is to improve the performance of the large-scale in-silico screening process in terms of preparation and execution time. The biological results of this activity can also enable medical chemists to better respond to instant and large-scale threats if the mutation of the H5N1 virus happens.

In parallel with exercising the enhanced workflow based on the WISDOM platform, a light-weight framework called DIANE (http://cern.ch/diane) was also taking part of the data challenge to provide an alternative scenario for handling massive molecular dockings on the Grid. Taking advantage of the DIANE framework, we demonstrate that the effort of application development, deployment and execution can be reduced. Moreover, stable throughput and efficient resource utilization can also be effortlessly achieved by adopting DIANE’s model of job control on distributed environments.

EGEE data challenge fighting against avian flu

H5 - The bird flu virus is named H5N1, H5 and N1 correspond to the name of proteins (hemagglutinin and Neuraminidase) on the virus surface.
- Hemagglutinin plays a major role in the virus multiplication
- Present drugs such as Tamiflu inhibit the action of Neuraminidase and stop the virus proliferation
- The N1 protein is known to evolve into variatns if it comes under drug stress
- To free up medicinal chemist’s time to better respond to instant and large-scale threats, a large scale in-silico screening was set for initial investment of the design of new drug

Problem size
- ~ 8 predicted possible variants of Influenza A neuraminidases
- around 300 K compounds from ZINC database
- A chemical combination library
- Computing challenge (a rough measurement based on Xeon 2.8 GHz)
- Each docking requires ~ 30 min CPU time
- Required computing power in total over 100 CPU years
- Storage requirement
- Each docking produces results with the size of 130 KByte
- Required storage space in total ~ 600 GByte (with 1 back-up)

To speed-up and reduce the cost to develop new drugs, high-throughput screening is demanded
- The WISDOM platform was used to reproduce high-throughput molecular docking
- 1/8 of the data challenge was taken by DIANE with the aspect of having an efficient and interactive control of the molecular dockings on the Grid

Distributing AutoDock jobs using DIANE

- DIANE – Distributed Analysis Environment
- A lightweight framework for parallel scientific applications in master-worker model
- Ideal for applications without communication between parallel tasks (e.g. for most of the Bioinformatics applications in analyzing a large number of independent datasets)
- The framework takes care of all synchronization, communication and workflow management details on behalf of the application
- Small effort for application development and deployment
- ~ 500 lines of python codes to distribute AutoDock on the Grid
- Flexible failure recovery mechanism
- Users can run multiple jobs, grid vs. application specific failure recovery
- Develop once, run in any environment

Application Oriented User Interface

DIANE Framework

Minimized effort of application development

Start the data challenge

```
> diane.stanier@diane.demo:autodock.job -ganga -n 1200 chgp -t 4 -n 1 -l ctn
```

Request more CPUs (DIANE workers)

```
> diane.stanier@diane.demo:autodock.job -ganga -n 2000 chgp -t 4 -n 1 -l ctn
```

Intuitive interface for job execution

Statsitics of the DIANE activity in DC2

- 280 DIANE worker agents were submitted as
- ~12 % failures related to application errors
- > 2000 CPUs used for 4 weeks
- ~ 12 % failures related to application errors
- > 2000 CPUs used for 4 weeks

Credit

ARDA : Jakub Moscicki
EGEE : Y. Legré
AuverGrid : E. Medernach
TWGrid : H.C. Lee, H. Y. Chen

Contact point: Y.T. Wu

H.C. Lee, M. Reichstadt
N. Jacq

Users (deputy)
J. Salzmann (N. Jacq)
M. Reichstadt (E. Medernach)
L. Y. H. C. Lee
I. Merelli, C. Arlandini (L. Milanesi)
J. Montagnat (Y. Gianardi)
R. Moillon (C. Blanchat)
I. Blanque (O. Segheli)
D. Garcia

Efficient handling of Large Scale in-silico Screening Using DIANE

Efficient handling of Large Scale in-silico Screening Using DIANE

Enable grids for E-sciencE

Enabling Grids for E-sciencE

In-silico Screening Using DIANE

Efficient handling of Large Scale