Contribution ID: **49** Type: **not specified**

Cloud, Ready for Bioinformatics ?

Wednesday, 28 March 2012 12:05 (25 minutes)

Description of the Work

The overall strategy for increasing the number and variety of users of the StratusLab infrastructure is to create a "virtuous cycle", starting with porting of a use case and publicizing it to generate more interest. A set of preliminary use cases, which will be the initial focus of the StratusLab porting activities, has been identified. Among seven use cases in total, two of them are related to bioinformatics and will be the focus of this presentation.

The first study, «Bioinformatics Services», is focused on creating bioinformatics appliances containing applications related to sequence analysis that scientists and engineers can deploy on demand. Indeed, several experimental technologies in the Biology area have been improved to such a degree that obtaining data is easy, causing a deluge of data for the bioinformatics community. The challenge is now to analyze efficiently these data with the relevant applications. Representative bioinformatics algorithms are for example BLAST, FastA or ClustalW that are data-intensive, processing gigabytes of data stored in flat-file databases like UNIPROT, EMBL or PDBseq via a shared filesystem, whereas other ones like Abyss, BWA, or Ray are CPU- and memory-intensive.

The second usecase, «TOSCANI: TOwards StruCtural AssignmeNt Improvement» is a project in collaboration between CNRS and Institut Pasteur to improve the determination of protein structures based on Nuclear Magnetic Resonance (NMR) information. This concerns the scientific disciplines around (i) molecular and structural biology (determination of biomolecular structures up to atomic resolution) and (ii) bioinformatics, which includes the ensemble of computer algorithms for treating the data from biological systems.

First releases of the required bioinformatics appliances has been developed and made available to the scientists via the StratusLab Marketplace.

Conclusions

First releases of the required bioinformatics appliances has been developed and made available for deployment on the project cloud infrastructures by the scientists via the StratusLab Marketplace. The representative bioinformatics tools has been selected and integrated in the virtual machine to set up StratusLab compliant appliances. These appliances have been made publicly available and are kept up-to-date. The StratusLab distribution provides all of the cloud services necessary to implement these use cases. In particular, they intensively use the image creation and management services as well as the data management facilities. The TOSCANI case will in addition requires a service manager for the deployment and management of the entire virtual infrastucture as a service.

Next steps will be to study a simple way of integrating new tools on-demand and to evaluate the usage by the community in terms of resources consumption, flexibility and elasticity brought by the cloud.

Impact

The adoption of clouds for bioinformatics applications will be strongly correlated to the capability of cloud infrastructures to provide ease-of-use and access to reference biological databases and common bioinformatics tools.
In the generic «Bioinformatics services» case and in extension to an overall perspective, cloud technologies provide scientists with the flexibility to deploy bioinformatics applications on different virtual machines. But clouds have to be connected with the existing public bioinformatics infrastructures. In that sense a cloud infrastructure should provide scientists with bioinformatics appliances to deploy on academic or commercial datacenters, or on their own computer or private cloud. Cloud deployment should also make the cloud infrastructure tightly connected to the storage of the the biological data. The interface of a cloud infratructure should ease the procedure of access by using the community's existing authentication methods (for example, single sign-on across portals and web services with Shibboleth technology). And of course providing access to a cloud will help bioinformaticians to build and to deploy single machines, clusters, or web service infrastructures to run a complete analysis pipeline.
About the TOSCANI study, significant increase in the number of calculated protein conformations improves the statistics on the NMR conformations and can help to overcome the ambiguity bottleneck. The large computing power required for this is concentrated in the simulated annealing procedure with the CNS software. Thus the elasticity of the cloud could be an advantage by waking the CNS VMs only during the simulation periods and putting them in sleep mode the rest of the time. Because of these large computational needs, an NMR laboratory not specially involved in bioinformatics developments will not invest in building a cluster of about 100 nodes to be able to run NMR structure calculations for example with ARIA.

URL

www.stratuslab.eu

Overview (For the conference guide)

The overall strategy for increasing the users community of the StratusLab infrastructure is to create a "virtuous cycle", starting with porting selected use cases to demonstrate the added value, and two of them are related to bioinformatics. The first study, «Bioinformatics Services», is focused on creating bioinformatics appliances providing tools and data related to sequence analysis that scientists and engineers can deploy on demand. The second usecase, «TOSCANI: TOwards StruCtural AssignmeNt Improvement» is a project in collaboration between CNRS and Institut Pasteur to improve the determination of protein structures based on Nuclear Magnetic Resonance (NMR) information. These usecases demonstrate the ease-of-use of use of an cloud infrastructure that stays connected with the existing public bioinformatics resources. First releases of the required bioinformatics appliances has been developed and made available to the scientists via the StratusLab Marketplace.

Primary authors: Dr LOOMIS, Charles (LAL, CNRS UMR8607); Dr BLANCHET, Christophe (IDB IBCP, CNRS FR3302); Mr GAUTHEY, Clément (IDB IBCP, CNRS FR3302)

Presenter: Dr BLANCHET, Christophe (IDB IBCP, CNRS FR3302)

Session Classification: Clouds: Users