Cloud BioLinux: using Amazon's cloud platform to provide pre-configured and on-demand high performance computing for bioinformatics.

BOSC 2010

Konstantinos (Ntino) Krampis
Bioinformatics Engineer
J. Craig Venter Institute

agbiotec@gmail.com
twitter: agbiotec

Introduction

While some mainstream tools such as the BLAST sequence comparison tool for example are accessible online through NCBI's website, most software packages need to be downloaded and configured by each individual researcher. Commonly-used bioinformatics tools are hard to build and maintain, usually available as source code or requiring numerous software library dependencies to choose from.This creates a burden for scientists, requiring them to provision the infrastructure and technical expertise, in order to use software developed as part of funded research. In addition, many bioinformatics workflows involve large datasets which require high performance computing setups for the analysis to be completed.

A tool developed specifically for alleviating these computational bottlenecks for genomic research is Cloud BioLinux (http://www.jcvi.org/cms/research/projects/jcvi-cloud-biolinux/overview/), a virtual machine publicly available on Amazon's EC2 platform. Cloud BioLinux provides software tools to the bioinformatics community following a Science as a Service (ScaaS) computing model. This is similar to the Software as a Service (SaaS) model, where providers offer pre-installed and configured applications, usually residing on a remote server and accessible via networking. With Cloud BioLinux as a base, we can provide pre-configured and on-demand high performance computing for the bioinformatics community.

Cloud BioLinux was based on Bio-Linux 5.0 (http://envgen.nox.ac.uk/tools/bio-linux), a linux distribution for desktop computers, which provides more than 100 bioinformatics packages pre-configured through an Ubuntu-Debian repository. In addition, this virtual machine includes the Celera Whole Genome Shotgun Assembler (http://wgs-assembler.sourceforge.net/), and the parallel computing implementation of the BLAST sequence comparison tool (http://www.mpiblast.org). Via scripts implemented within Cloud BioLinux, push-button virtual parallel computing clusters can be created on Amazon EC2, for performing large scale sequence analysis.

Hackathon goals

A community of bioinformatics developers has been already established around Cloud Biolinux, and members have agreed for it to be a reference virtual machine (http://lists.cloudbiolinux.com/pipermail/community/2009-December/000001.html), for porting bioinformatics software on cloud computing platforms. During this workshop, we plan to package additional software which is commonly used in bioinformatic analysis pipelines, in the Cloud Biolinux virtual machine. In addition, we plan to provide a tutorial for researchers interested in using Cloud Biolinux as part of their analysis pipelines.

In parallel to this goal, we will work to implement a configuration and software management framework for bioinformatics applications that are currently bundled with Cloud BioLinux, but also for those that will be included in future versions. This will be achieved through setting-up an Ubuntu-Debian Linux repository, a simple web-interface, and scripts on the back-end to drive and keep track of the installed packages. The framework's foundation will be either Puppet or Chef (http://reductivelabs.com/), which are currently used by the IT industry for maintaining large numbers of server configurations. Once implemented, this framework will provide flexibility to researchers to customize the virtual machine for the needs of their individual groups. In addition, it will help managing the growing software collection in the Cloud BioLinux virtual machines, but also allow members of the community to "fork" it and keep track of package updates. Finally, our framework will be widely applicable to managing software installations, in other virtual machines other than Cloud Biolinux running on Amazon's EC2.

Conclusion

In the Cloud Biolinux project we combined the convenience of SaaS for end-users with the power of cloud computing, in order to provide pre-configured and on-demand high performance computing for bioinformatics. Through virtual infrastructures and cloud computing, access to computational resources can be democratized, especially for scientists based at smaller research institutions.