Enhancing Grid Infrastructures with
Virtualization and Cloud Technologies

Creation of Virtual Appliances for Bioinformatics Community

Milestone MS3 (V1.0)
14 March 2011

Abstract

IBCP has created two customized machine images for the bioinformatics community: "biological databases repository" and "bioinformatics compute node". The "biodata repo" VM aims to provide users with access from any cloud node to international reference databases recording biological resources such as protein or gene sequences and associated data, protein structures, or complete genomes. This appliance acts as an proxy between the internet where all the reference databases are published and the cloud internal virtual nodes that will compute the bioinformatics analyses. The "biocompute node" VM has pre-installed bioinformatics software such as ClustalW, BLAST, FastA and SSearch. Because these methods require access to reference data for processing, this appliance is linked via an NFS mount to the "biodata repo" appliance.

StratusLab is co-funded by the European Community's Seventh Framework Programme (Capacities) Grant Agreement INFSO-RI-261552.

The information contained in this document represents the views of the copyright holders as of the date such views are published.

Contributors

Name	Partner	Sections
Christophe Blanchet	CNRS/IBCP	All

Document History

Version	Date	Comment
0.1	10 Mar. 2011	First draft for internal review.
1.0	14 Mar. 2011	Final draft after internal review.

1 Introduction

Today, the bioinformatics community is facing a deluge of data. Several experimental technologies have been improved in such a way that obtaining data is easy: for example from the study of a single gene/protein to a whole genome or family of proteins, from a single metabolic pathway to systems biology. The challenge is to be able to analyze these data with the relevant applications. For example, sequencing a whole genome has became usual with the new technologies called Next Generation Sequencing (NGS). Many projects are working on the genome sequence of different organisms, thus continuously providing new sequences for analysis. Such analyses are very demanding in terms of computation at the primary level for the assembly of the short-reads and at the secondary level in terms of sequence analysis for the comparison of the new data to the reference databases. Algorithms like BLAST, FastA and SSearch are used intensively for that analysis and usually classified as data-intensive. They are processing in tenths of seconds gigabytes of data stored in flat-file databases like UNIPROT, EMBL or PDBseq on a shared filesystem.

IBCP has created two customized machine images for the bioinformatics community: a “biological databases repository” and a “bioinformatics compute node”.

2 Bioinformatics appliances

Cloud-based resources can bring increased flexibility to bioinformatics users using applications with specific software requirements for their scientific experiments. The ability to predefine specific and well-suited virtual machines fulfilling the requirements of, sometimes challenging, bioinformatics applications, and to make them available from central repositories ready for deployment on academic clouds, would satisfy the demanding bioinformatics user. An obvious example is the ability to run applications with contradictory requirements in terms of RAM or software dependencies on separate virtual machines. Other ones related to scientific constraints are to provide scientists and engineers with predefined appliances for example, for different requests like protein sequence analysis that require a lot of single-CPU machines but connected to up-to-date reference biological databases. Another example of genome assembling requires in most cases few CPUs but needs to be linked to a large shared memory of hundreds of gigabytes.

The usage of cloud for bioinformatics has to be connected with public bioinformatics infrastructures like the RENABI GRISBI at the French level or the ELIXIR infrastructure at the European level. The adoption of clouds for bioinformatics applications will then be strongly correlated to the capability of cloud infrastructures to provide access to reference biological data and applications.

2.1 Biological databases repository

Regarding the biological data, bioinformaticians must have access from any cloud node to international reference databases recording biological resources such as protein or gene sequences and associated data, protein structures or complete genomes. These databases are annually referenced in an annual "Database" issue of the scientific journal Nucleic Acids Research[1]. The 2011 edition lists 1330 carefully selected molecular biology databases.

We have built a virtual machine on StratusLab for that purpose. This machine requires significant storage and high-availability features as it will be the reference endpoint in terms of biological data for all the bioinformatics users on a particular cloud. We have used the "CentOS 5.5" base image from the StratusLab repository and matched it with a large VM type (4 CPUs, 2048 MB RAM). This appliance acts as an proxy between the internet where all the reference databases are published

[1] http://www.oxfordjournals.org/nar/database/c

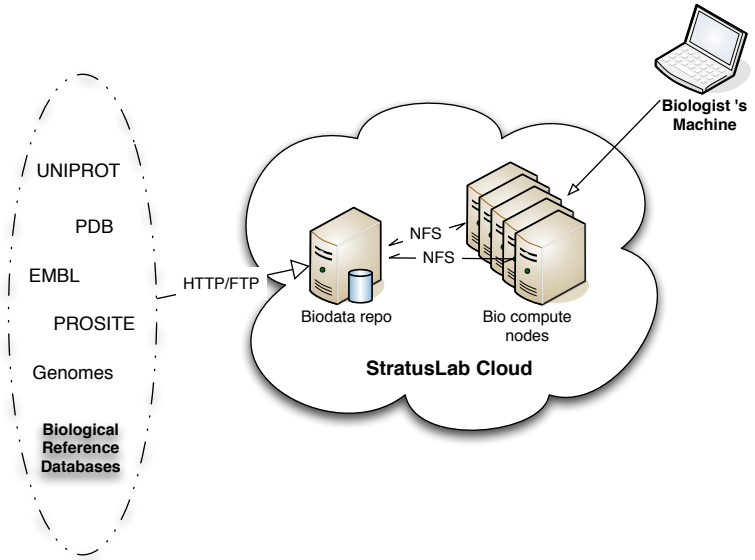

Figure 2.1: *Bioinformatics infrastructure on a StratusLab cloud*

and the cloud internal virtual nodes that will compute the bioinformatics analyses (see Figure 2.1).

To import and maintain the required biological databases, we have used the BioMaJ system developed in France by several RENABI platforms[2]. BioMaj has software dependancies mainly on perl, ant, java, httpd, tomcat6 and mysql-server. Within the BioMaj system, we filled a properties file with the related parameters for each database we want to install and keep updated. Once the property files are installed for the chosen databases, BioMaj regularly checks if some bases need to be updated and stores the data in files organized from a root directory '/biodb'. We have also configured a read-only NFS export of this root 'biodb' to all the bioinformatics computing machines of the cloud. For those reasons, it is very important that this virtual appliance has high-availability feature and is being kept running even if the StratusLab physical node crashes.

Although NFS sharing may not be efficient at a large scale, it is required by the bioinformatics applications we are running on the bioinformatics compute node. Indeed, some bioinformatics tools like BLAST or FastA require a standard POSIX local access to the flat-file databases used as references for the computational analysis. A promising perspective would be to have an EBS-like volume on the StratusLab cloud that the "biological databases repository" instance will mount in a read-write mode to install and update the databases. And that the "bioinformatics compute" node instances will mount in a read-only mode to make the bioinformatics tools connected to the reference data. Having an EBS-like system will also help to solve the demand of such a cental repository in term of storage size. For

[2]http://biomaj.genouest.org

example, in IBCP we have such a databases repository in production with a 5TB storage for the DB, almost full.

2.2 Bioinformatics compute node

Distributing the computation is also an important requirement because bioinformatics applications could require very different resources depending on the analysis to perform: multiple alignments of sequences, genome assembling or intensive protein sequence comparison. Biologists and bioinformaticians are combining regularly multiple software packages to analyze their data, which they mainly access them from the command line for their intensive processes.

Regarding the computations and the virtual machines, the main requirements are related to satisfying the software dependencies and the very different behavior of the biological applications in terms of CPU and memory. Some applications only require one CPU but with a lot of memory (96 or 128MB) whereas others require lot of CPUs that are accessed through MPI mechanism. We have built a virtual machine with pre-installed bioinformatics software. For that we have used the "CentOS 5.5" base image from the StratusLab repository. To install the required bioinformatics software we used scripts, called 'bioapps', that we had developed. This system uses an URL from where the source code can be downloaded and a location where to install the compiled binary on the machine. Because these tools require access to reference data to process their analyses, this bioinformatics compute appliance is linked to the previously-defined biological databases appliance, and require to mount the exported volumes containing the biological data.

We have predefined a bioinformatics appliance with software such as ClustalW, BLAST, FastA and SSearch, which was presented during one of the end-of-sprint demonstration meetings. Yet users should connect and run the application by hand, but we are planning to added Web interfaces. That could be a local Web portal where the user connect to input his data and run the tool. Or that could be a Web service interface (with SOAP or RESTful endpoint) that the user could integrate to its standard bioinformatics workflows.