StratusLab

Enhancing Grid Infrastructures with Virtualization and Cloud Technologies

Virtual Appliances for the Bioinformatics Community

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. 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. Algorithms like BLAST, FastA or ClustalW are used intensively for that analysis and usually classified as data-intensive. They are processing gigabytes of data stored in flat-file databases like UNIPROT, EMBL or PDBseq on a shared filesystem. To give an insight to this challenge, StratusLab has built two virtual appliances, "Biological databases repository" and "Bioinformatics compute node", to provide bioinformaticians with a repository appliance maintaining up-to-date international reference databases, then made available through shared filesystem in destination to bioinformatics cloud nodes with pre-installed bioinformatics software.

The usage of cloud for bioinformatics has to be connected with public bioinformatics infrastructures like the French **Bioinformatics Network RENABI** (www.renabi.fr) and especially its **grid infrastructure GRISBI** (www.grisbio.fr). 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 data and applications. In that sense, StratusLab is collaborating with RENABI to help solving the requirements from the Bioinformatics community.

Biological databases repository appliance

Bioinformaticians need access from any compute 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. The 2011 edition lists 1330 carefully selected molecular biology databases.

We have built a virtual appliance that acts as a proxy between the internet where all the reference databases are published and the cloud instances that will compute the bioinformatics analyses. To import and maintain the required biological databases, we have used the BioMaJ system developed in France by the RENABI network. Once the property files are installed for the selected 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 needed by some bioinformatics applications like BLAST or FastA that require a standard POSIX local access to the flat-file databases used as reference for the computational analysis. A promising perspective would be to have an EBS-like volume on the StratusLab cloud that the "biological databases repositor" instance will mount in a read-write mode to install and update the databases. And that the "bioinformatics comput" 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 central repository by providing efficiently terabytes of shared storage.

Bioinformatics compute node appliance

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. They used these software for their intensive processes from Web portals, or through with shell commands or scrips written in interpreted languages.

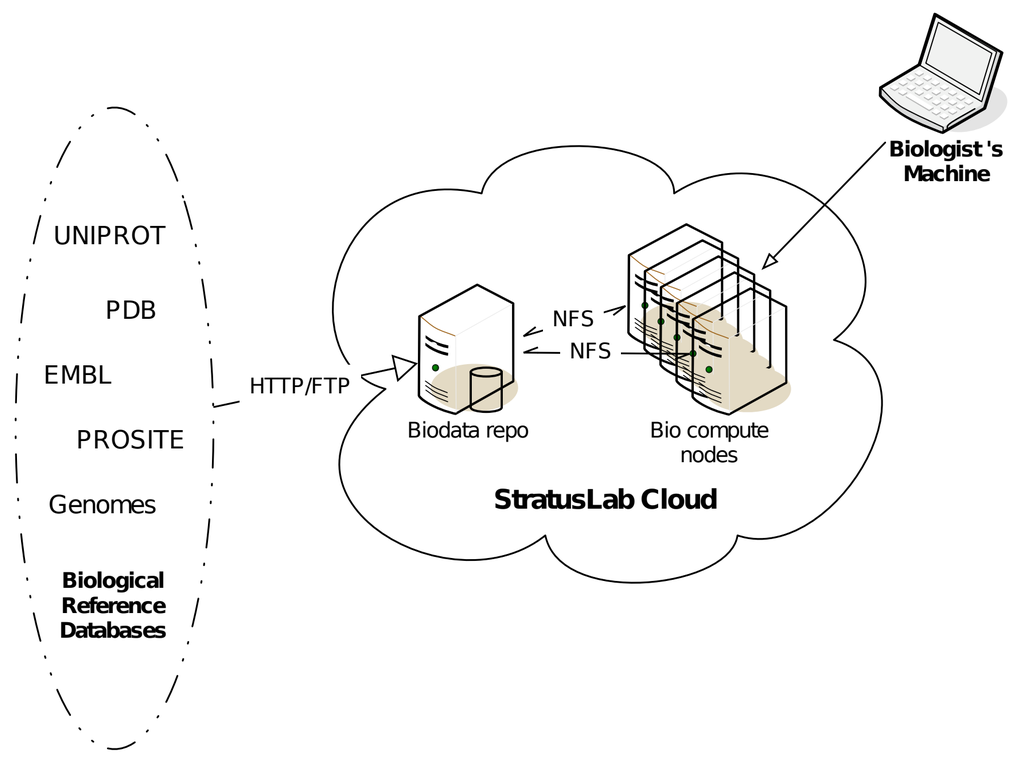
Regarding the computations and the virtual machines, the main requirements are related to satisfying the software dependencies and the very different behavior of the biological applications in terms of CPU and memory. Some applications only require one CPU but with a lot of memory (96 or 128MB) whereas others require lot of CPUs that are accessed through MPI mechanism. We have built a virtual machine with pre-installed bioinformatics software. To install the required bioinformatics software we used a script system, called `bioapps', that we had developed. This tool download the application package from the reference site and install the compiled binary on the machine. Because the bioinformatics applications require access to reference data to process their analyses, this bioinformatics compute appliance is linked to the biological databases repository 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. 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 Web service interfaces (with SOAP or RESTful endpoint) that the user could integrate to its standard bioinformatics workflows.

Bioinformatics requirements regarding the cloud

Flexibility to use bioinformatics applications with specific software requirements. Cloud for Bioinformatics has to be connected with public Bioinformatics infrastructures especially the biological databases.

PaaS
Provide scientists with Bioinformatics appliances on academic or commercial datacenters, e.g. RENABI, or on their own computer/private cloud.
Missions of bioinformatics centres switch from providing services to providing virtual appliances.

IaaS
Bioinformaticians can pre-define and deploy cluster, or WS servers.
infrastructure Biologists can deploy the required infrastructure according to their analysis pipeline.



Website: http://stratuslab.eu
Support: support@stratuslab.eu
Contact: contact@stratuslab.eu

Centre National de la Recherche Scientifique *France*
Universidad Complutense de Madrid *Spain*
Greek Research and Technology Network S. A. *Greece*
SixSq Sàrl *Switzerland*
Telefonica Investigacion y Desarrollo S.A. *Spain*
Trinity College Dublin *Ireland*

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