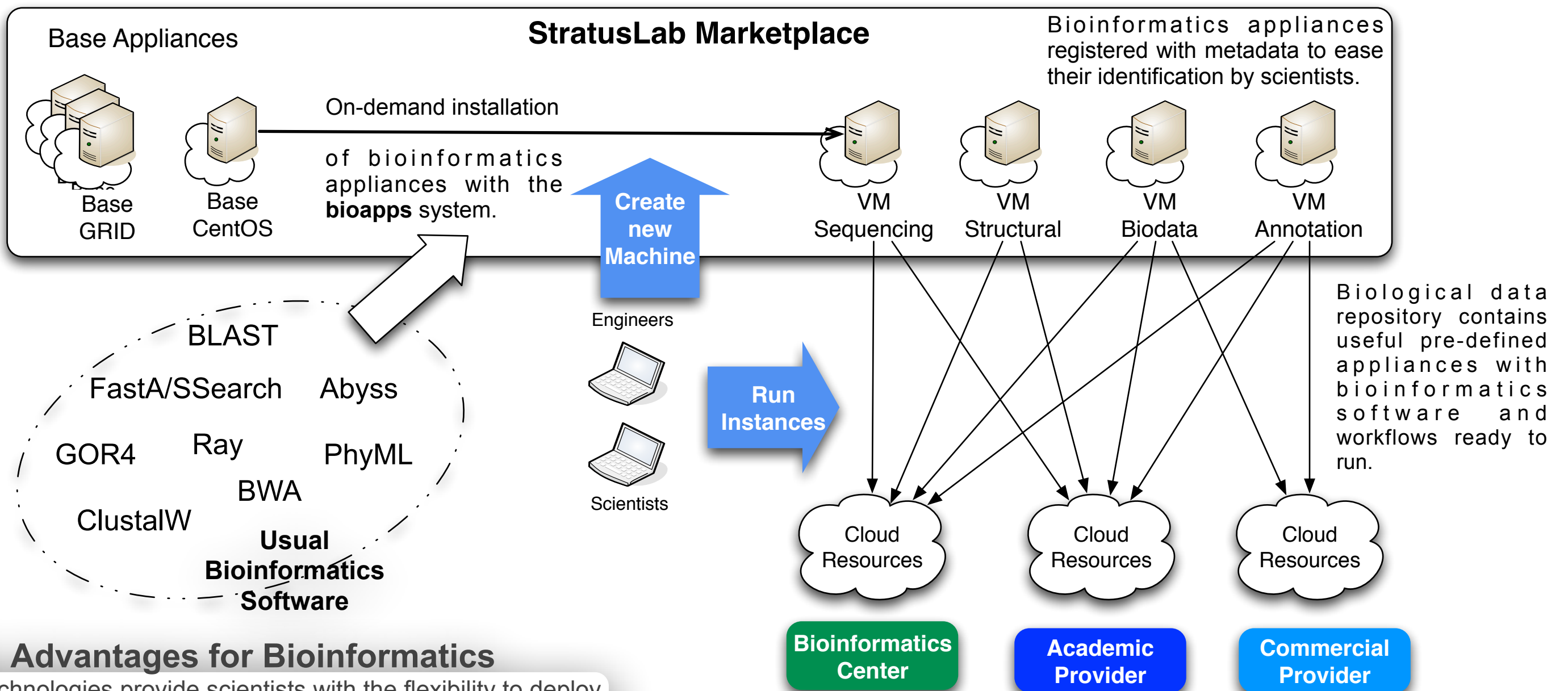


Several experimental technologies have been improved to such a degree that obtaining data is easy, causing a deluge of data for the Bioinformatics community. The challenge is to be able to analyze these data with the relevant applications, for example, sequencing a whole genome obtained from Next Generation Sequencing (NGS) instrument. Many projects are working on the genome sequence of different organisms, continuously providing new sequences for analysis. Some bioinformatics algorithms like BLAST, FastA or ClustalW are used for that analysis and are usually classified as data-intensive, processing gigabytes of data stored in flat-file databases like UNIPROT, EMBL or PDBseq via a shared filesystem. Others like Abyss, BWA or Ray take the output sequences of sequencing machines and assemble them to get the complete sequence of the studied genome. In the context of the StratusLab project (EU-FP7, www.stratuslab.org), we have built two bioinformatics virtual appliances: a "Biological databases repository" and a "Bioinformatics compute node".



Cloud Advantages for Bioinformatics

Cloud technologies provide scientists with the flexibility to deploy bioinformatics applications on different virtual machines. But clouds have to be connected with public bioinformatics infrastructures, especially the network of international biological databases, to be useful.

Platform as a Service - PaaS

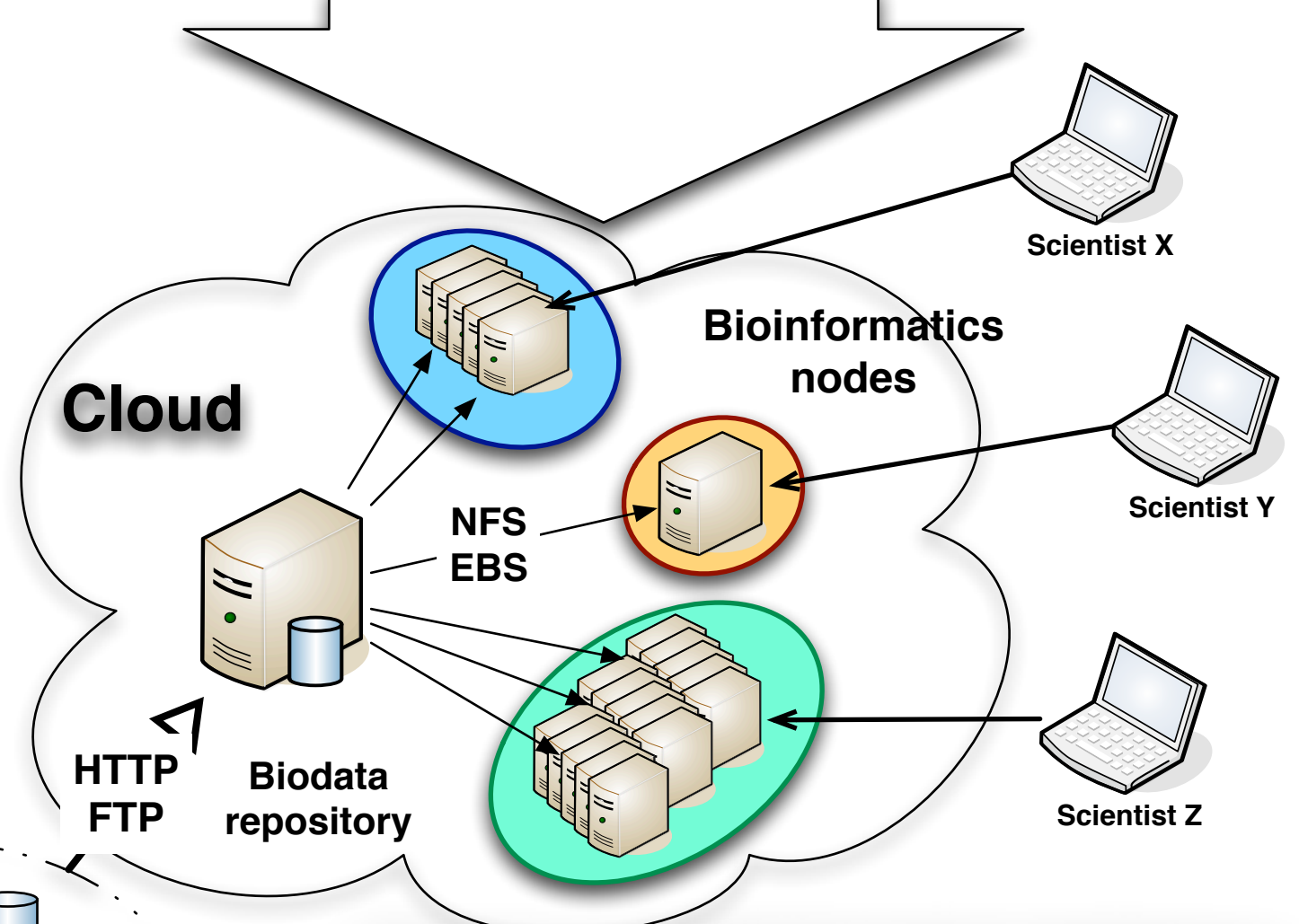
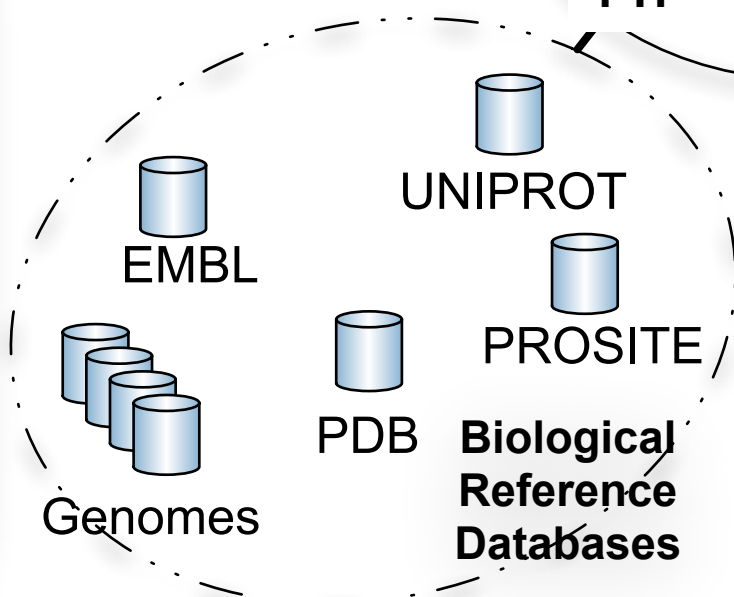
- Provide scientists with Bioinformatics appliances on academic or commercial data centers, e.g. RENABI, or on their own computer or private cloud
- The mission of bioinformatics centers switches from providing services to providing virtual appliances

Infrastructure as a Service - IaaS

- Bioinformaticians can pre-define and deploy clusters or web service infrastructures
- Biologists can deploy different virtual servers to run a complete analysis pipeline

Access to Reference Databases

Access from any compute node is a strong requirement. Such databases contain, for example, protein or gene sequences and associated data, protein structures or complete genomes. We have built a virtual appliance that acts as a proxy between the internet where all the reference databases are published and the cloud where the bioinformatics analyses will be computed. To import and maintain these databases, we use the BioMaJ system (biomaj.genouest.org). This virtual machine stores the data on a local disk (in the future on a cloud-persistent disk), and then exports it with NFS to all the bioinformatics computing machines of the cloud.



Ready-to-run Bioinformatics Appliances

Appliances containing common bioinformatics software is also an important requirement. We are building such appliances, keeping them up-to-date, and making them available from the StratusLab Marketplace. Scientists with a cloud login can then launch as many instances of these appliances as required by their analysis pipelines. Because bioinformatics applications require access to reference data to process their analyses, the compute node mounts the exported volumes containing the biological databases.

Bioinformatics Cloud Usage

Usage must 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 databases and common bioinformatics software. In that sense, StratusLab is collaborating with RENABI to help fulfill the requirements of the Bioinformatics community.