# Clouds in biomedical sciences
# Part III – clouds in biosciences

Vincent Breton

July 28th 2014

Enrico Fermi school of physics

# Session III: clouds in life sciences

- Generalities
- Deployment of life science applications on public clouds
- “De novo” deployment deployment of scientific applications on academic clouds
- Pilot jobs platform help hiding technical difficulties
  - examples

# Summary of grid adoption in life sciences

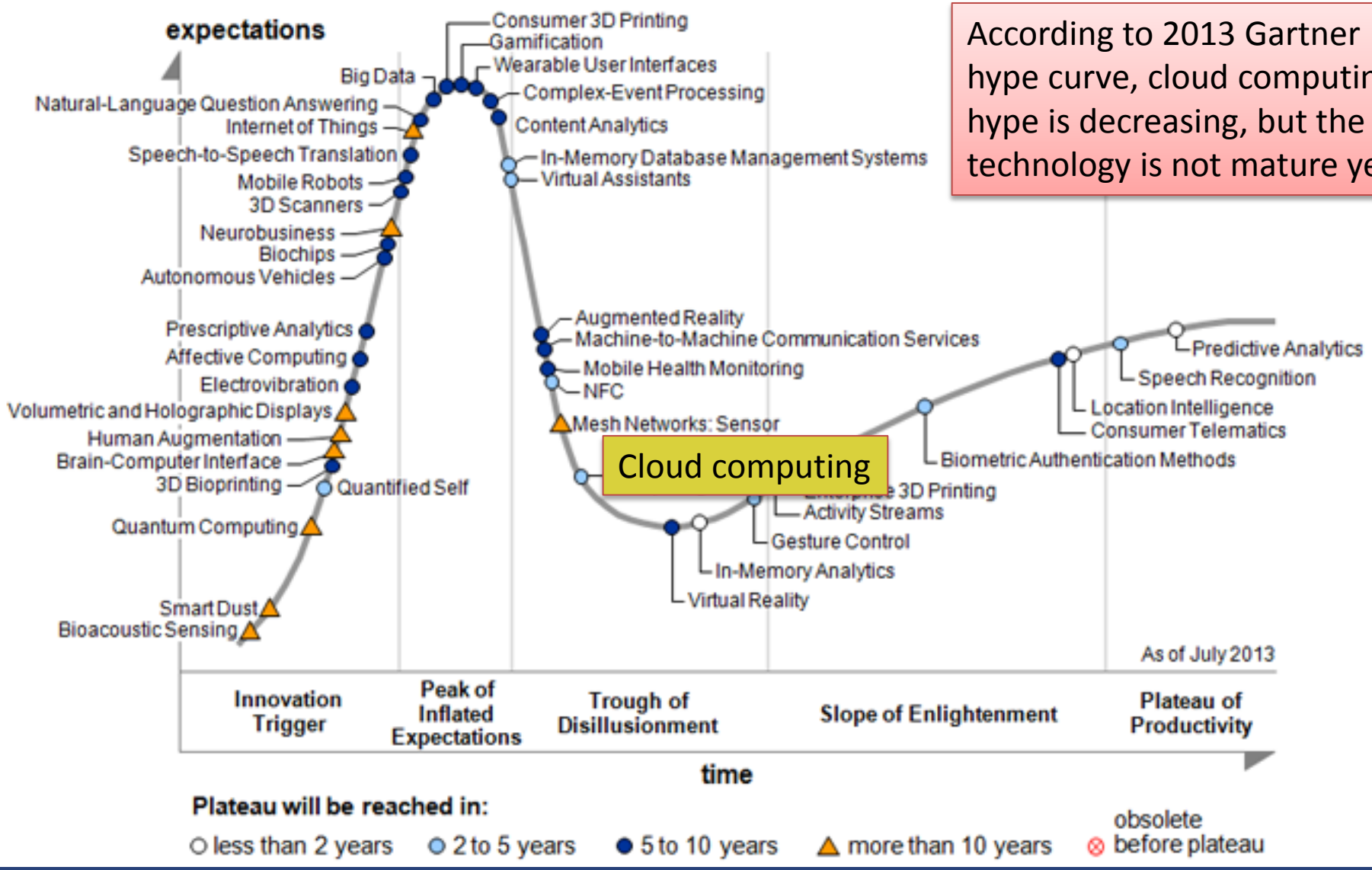
| Scientific subdiscipline | Achievements | Limitations |
|---|---|---|
| Structural biology | 100s of users through scientific gateways | Grid operational cost |
| Drug discovery | Large scale deployment of docking computations | IP issues have stopped adoption |
| Medical imaging (simulation) | 100s of users through scientific gateways | Grid operational cost |
| Neurosciences | Emergence of grid-enabled scientific gateways | Protection of medical data – grid operational cost |
| Molecular biology - bioinformatics | Limited adoption | Grid middleware OS – Data management – grid operational cost |

Cloud computing provides new opportunities (flexibility, reduced operational cost)

# The promises of cloud computing

- Public clouds
  - No cost to operate IT infrastructure: only pay what you use
  - Computing capacity on demand
    - Unbound resources
  - Flexibility to upload favorite Operating System
- Academic (private) clouds
  - Reduced cost to operate IT infrastructure (compared to grid)
  - Flexibility to upload favorite Operating System

# Where are we today?

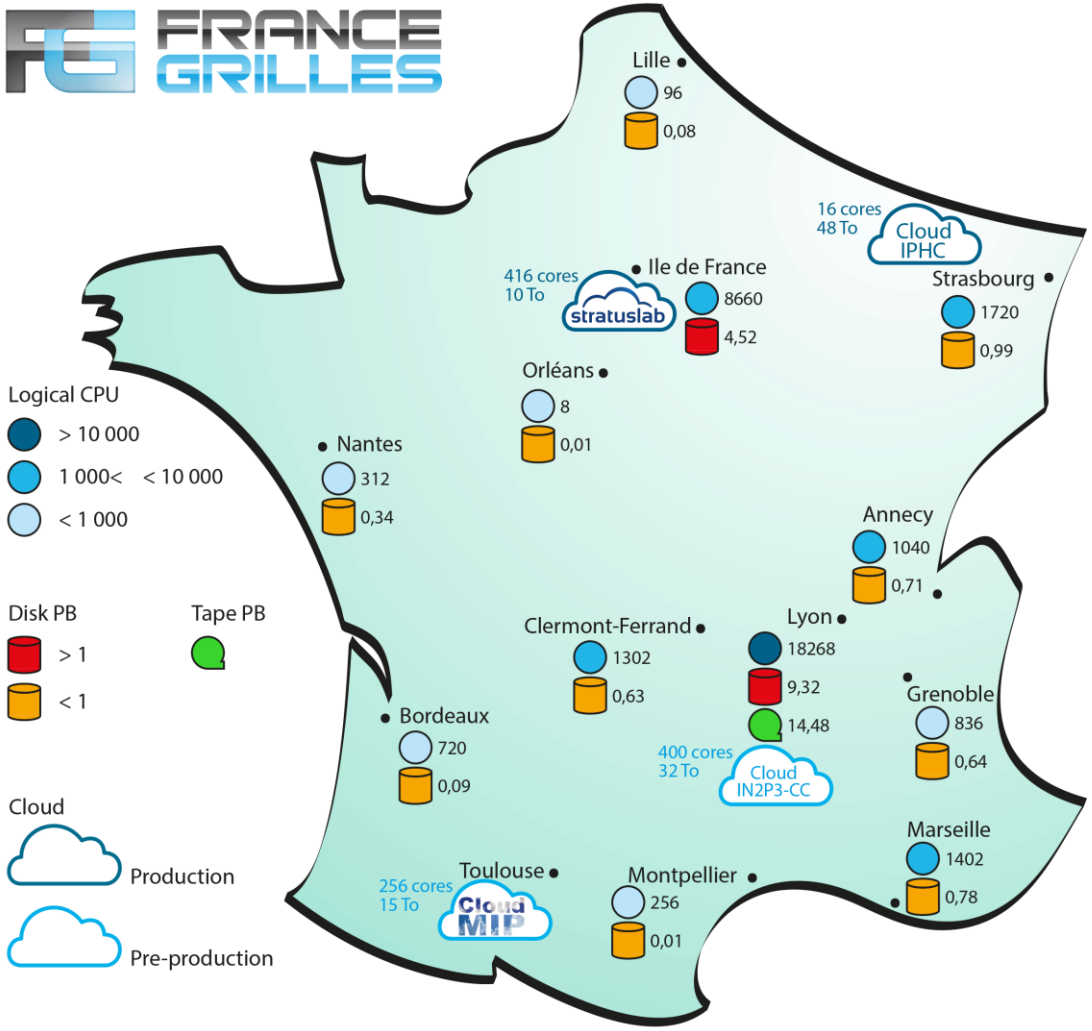



Cloud computing

According to 2013 Gartner hype curve, cloud computing hype is decreasing, but the technology is not mature yet

# The situation in France

- French state put all cloud money in industry
- Federation of academic clouds started in 2012
  - OpenNebula
  - StratusLab
  - OpenStack



# Adoption of clouds in the life sciences community in 2014 is very hard to assess

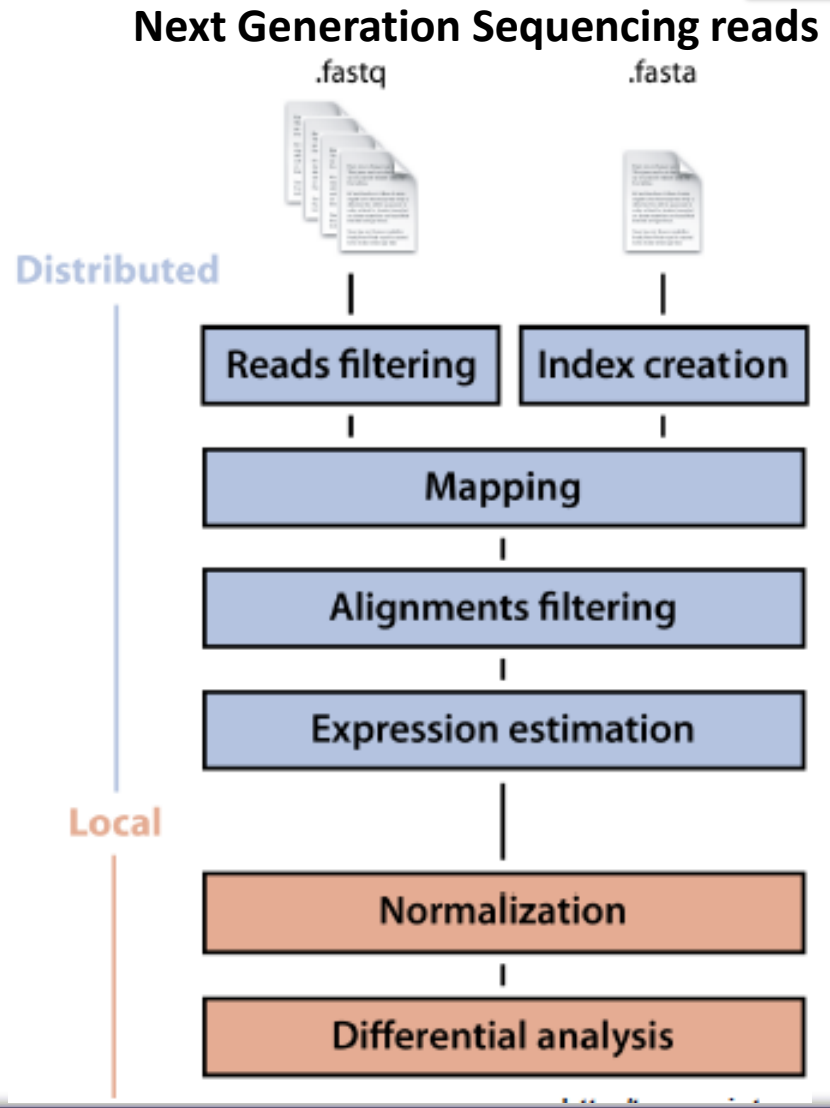
- Everything is now renamed cloud computing
  - Cluster computing
  - Grid computing
- Three scenarii:
  - Deployment of scientific applications on public clouds (Amazon)
  - De novo deployment of scientific applications on academic clouds
  - Migration to academic clouds of grid applications deployed using pilot agent platforms

# Deployment of life science applications on public clouds

- Only a few research groups are using public clouds in France
  - Academic Research funding model is hardly compatible with credit card payment for computing capacity
- Feedback is not very positive
  - Public clouds perceived as expensive compared to academic clusters/grids

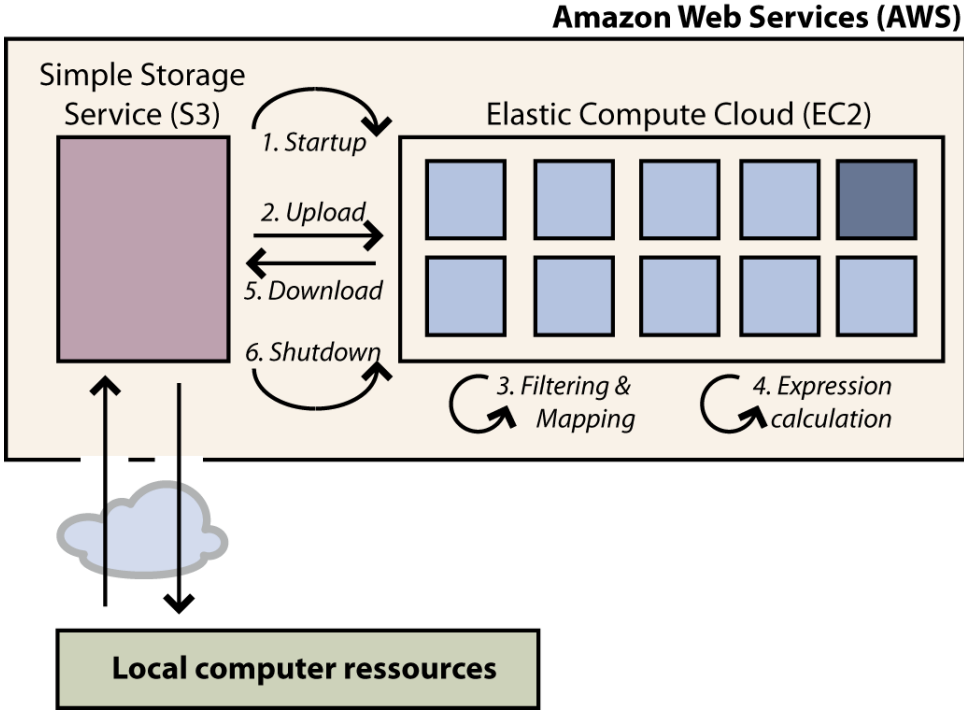
# Eoulsan experience on AWS (Amazon)

- Eoulsan is an analysis workflow of RNA-sequences
- Three steps:
  - Data upload (upload step)
  - Read mapping and filtering (filtermap step)
  - Transcript abundance estimation (expression step)
- Distributed calculations to speed up analysis
  - Parallelisation using Hadoop



# Eoulsan conclusion: Grid performances better than Amazon Web Service

- Comparison of Eoulsan running times (in minutes) between grid and Amazon cloud (AWS) for each analysis step
  - Human data
  - 888 Million reads corresponding to 88Gb data
- Conclusion: migration to EGI of the pipeline analysis




| | Upload | filtermap | expression | Total |
|---|---|---|---|---|
| Standalone | 154 | 1,146 | 4 | 1,304 |
| Grid | 53 | 388 | 2.5 | 467 |
| AWS | 80 | 810 | 64 | 1,120 |

# Some considerations on public cloud storage prices

- Google Drive offer (⇔ external hard disk): 1$ per TeraOctet per month [1]
- Storage offers on commercial clouds: ≈ 300K$/PO/yr
  - Amazon S3[2] and Google[3] almost equivalent: ≈ 30$ per TeraOctet per month
  - Additional cost: billing of requests and data transfers
    - Amazon S3: 0,1 $ per GOctet of data transfered from S3 to internet (100K$/PO)
    - Google: ≈ 0,2 $ per GOctet of data transfered from S3 to internet (200K$/PO)

[1]: valid for 300 Toctets and above
[2]: http://aws.amazon.com/fr/s3/pricing/
[3]: https://cloud.google.com/products/cloud-storage/#pricing

# De novo deployment of scientific applications on academic clouds

- Ecclesiastes 1:9* The thing that hath been, it *is that* which shall be; and that which is done *is* that which shall be done: and *there is* no new *thing* under the sun.

# Example: the e-Biothon initiative

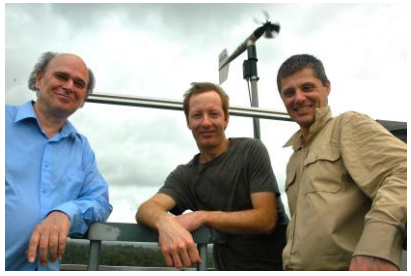

- Telethon: every year, fund raising by french media for French Muscular Distrophy Association (AFM)
- From Telethon to Decrypthon
  - Computing infrastructure (IBM)
  - Research projects (CNRS)
  - Human resources (AFM)
- From Decrypthon to E-Biothon

# E-Biothon: infrastructure

- 2 Blue Gene/P IBM racks with 200 TO storage
  - 2x1024 4-core nodes
  - up to 28 TFlops peak performance
- **SysFera-DS web access to computing resources**
- 2 modes:
  - Standard (MPI)
  - **HTC (1024 independent tasks in parallel)**

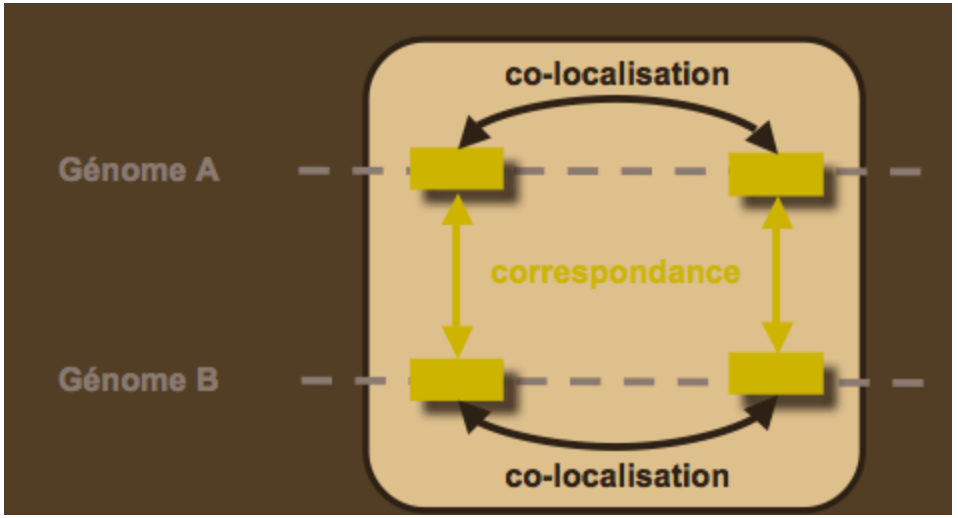
# E-Biothon vision is to offer a service to the user communities in life sciences

- 2013-2014: first 3 projects
  - Jean-François Gibrat et al, (MIGALE platform, INRA Jouy-en-Josas)
  - Olivier Gascuel, Stéphane Guindon et Vincent Lefort (CNRS Montpellier)
  - Yec'han Laizet, Philippe Chaumeil, Jean-Marc Frigerio, Stéphanie Mariette, Sophie Gerber, Alain Franc (INRA BioGeCo – Bordeaux)
- > 2014: open call for projects (IFB)



# Studying the synteny over a wide range of microbial genomes

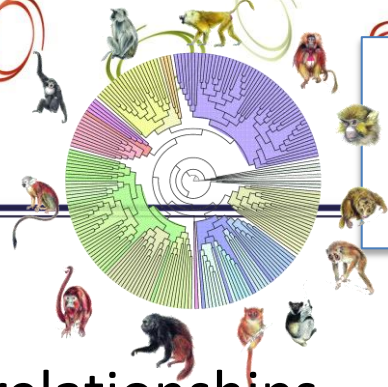
- Definition: similar blocks of genes in the same relative positions in the genome




- Interest: Study of synteny can show how the genome is cut and pasted in the course of evolution
- MIGALE team at INRA designed a pipeline analysis to compute synteny between 2 genomes and store it in a database
- **E-Biothon impact: change in scale - capacity to compute synteny between 2000 complete bacterial genomes (7 millions comparisons**)

Phylogeny.fr
Robust Phylogenetic Analysis For The Non-Specialist

# PhyML



IDGC
Institut des Grilles et du cloud du CNRS

Philogenetics is the study of evolutionary relationships among groups of organisms

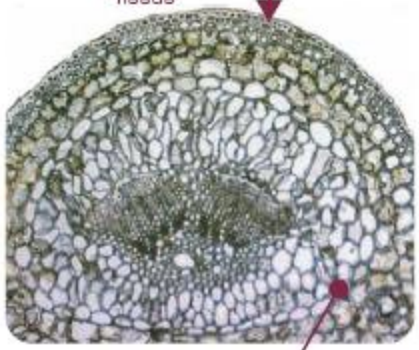
*PhyML* is a software that estimates maximum likelihood phylogenies from alignments of nucleotide or amino acid sequences

PhyML original publication in 2007 is the most cited in environment and ecology (> 6000 citations).

**e-Biothon impact: change in scale in the resources made available to PhyML users**

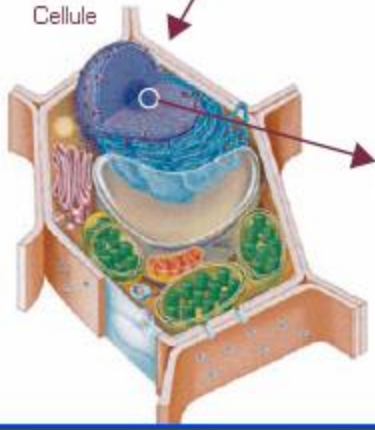
# Characterizing biodiversity

Tissus

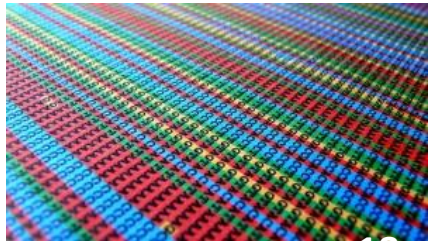


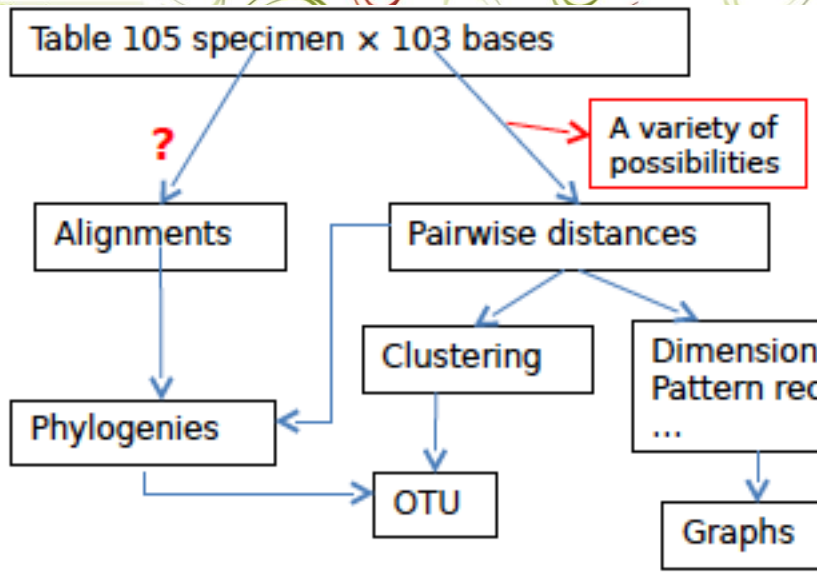
ACGTGTGCTAT ▶ Quercus petraea
ACGCGTGCTAT ▶ Quercus robur
ACGT--GCTAT ▶ Quercus pubescens
ACGCAGTCTAT ▶ Quercus cocinea

Cellule



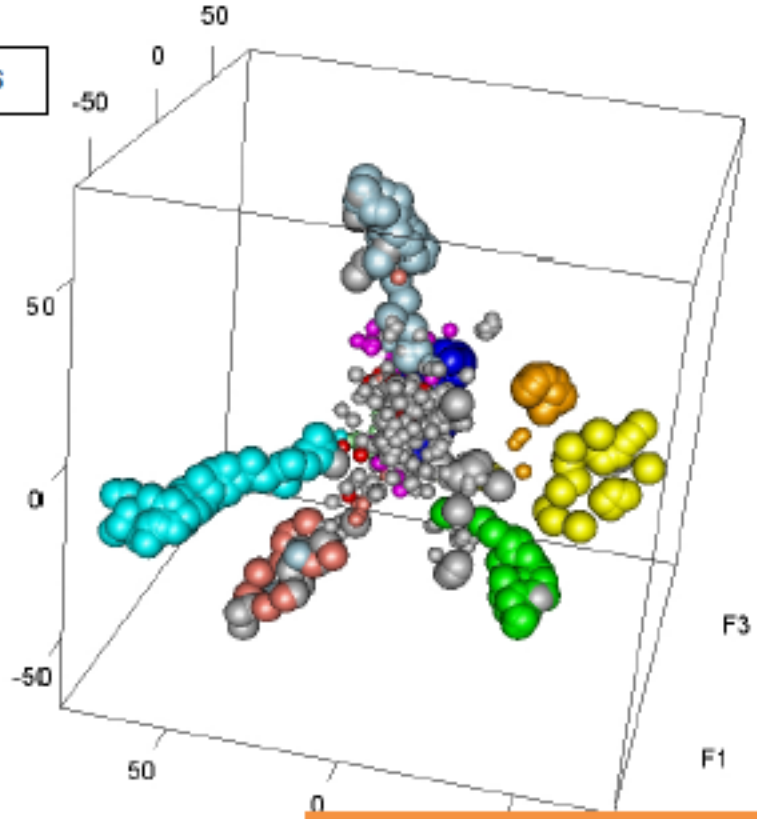
Extraction,
Amplification,
Séquençage
ADN






According to botanic theory, biodiversity is organized in species, genders, families, orders: is it confirmed in the distance between sequences?

blue -> Mimosoideae
---
lightblue -> Lecythidaceae
---
cyan -> Chrysobalanaceae
---
green -> Annonaceae
---
lightgreen -> Caesalpinioideae
---
yellow -> Myrtaceae
---
orange -> Elaeocarpaceae
---
magenta -> Apocynaceae
---
salmon -> Burseraceae
---
red -> Malvaceae
---




~ 1000 individus

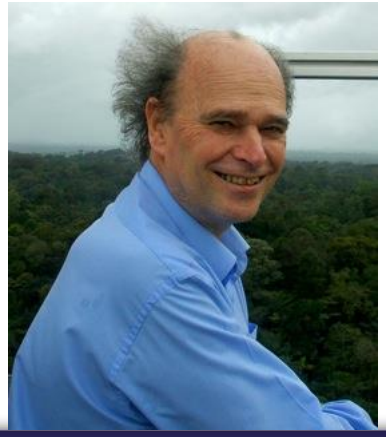
# Study of biodiversity in Guyane

16000 different tree species in amazonian forest (≈ 300 in Europe)

More biodiversity in 10000 $m^2$ of forest in French Guyana than in Europe

**E-Biothon added value**

- Change in scale (from local Mesocenter in Bordeaux)
- Millions of reads
- Exact distance computation without heuristics (alignement scores)





Credit: Alain Franc et Yec'hran Laizet

# Which global strategy for molecular biology ?

- Grid middleware and computing resources do not optimally fit the core needs of molecular biology
  - Genome assembly from Next Generation Sequencing raw data requires both RAM and large disk storage
  - Bioinformatics analysis requires much more flexibility than current grid infrastructures

# The french strategy for molecular biology

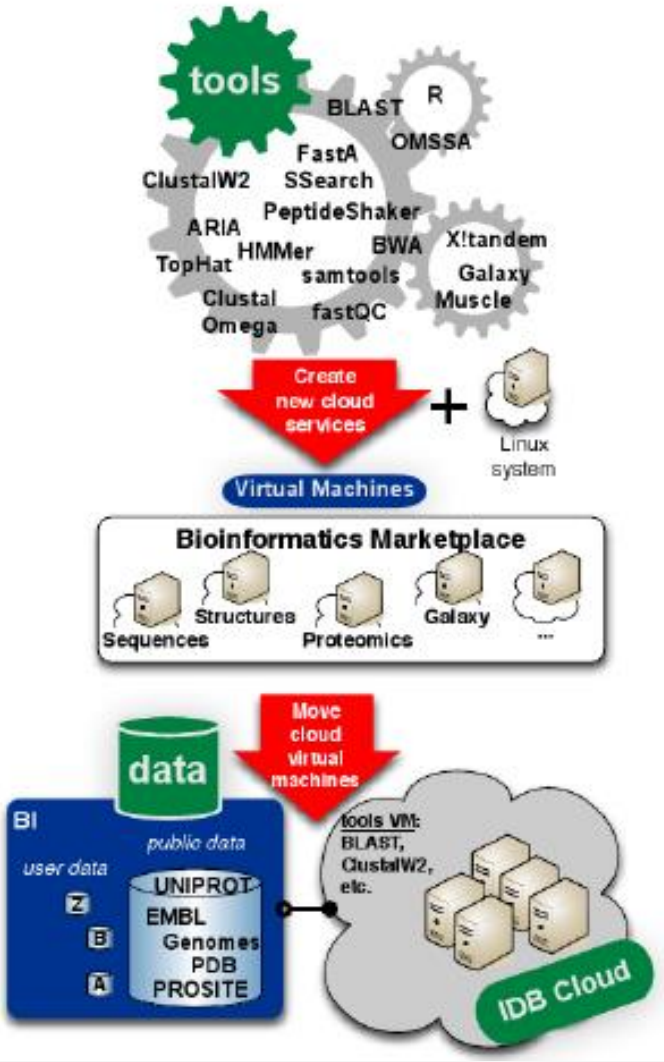
- France Genomique: an infrastructure to strengthen french capacities for High Throughput genomics
  - Central resource: HPC computing and storage resources @ TGCC (CEA)
- Institut Français de Bioinformatique: an infrastructure for the management and analysis of biological data
  - Central resource: academic cloud @ IDRIS
  - French node of ELIXIR, the European Research Infrastructure for Molecular Biology

# France Genomique @ TGCC

- Computing resources
  - 180 bi processors nodes (Intel Sandy Bridge E5-2680, 2.7 GHz, 8 cores) with 128 Go memory per node, equivalent to 2.880 cores (Bull)
  - 2 very large memory systems Bullx S6410 systems with 2 To memory
- Storage resources: 5 Po including 2 Po on disk
  - Hierarchical storage system Lustre + IBM HPSS

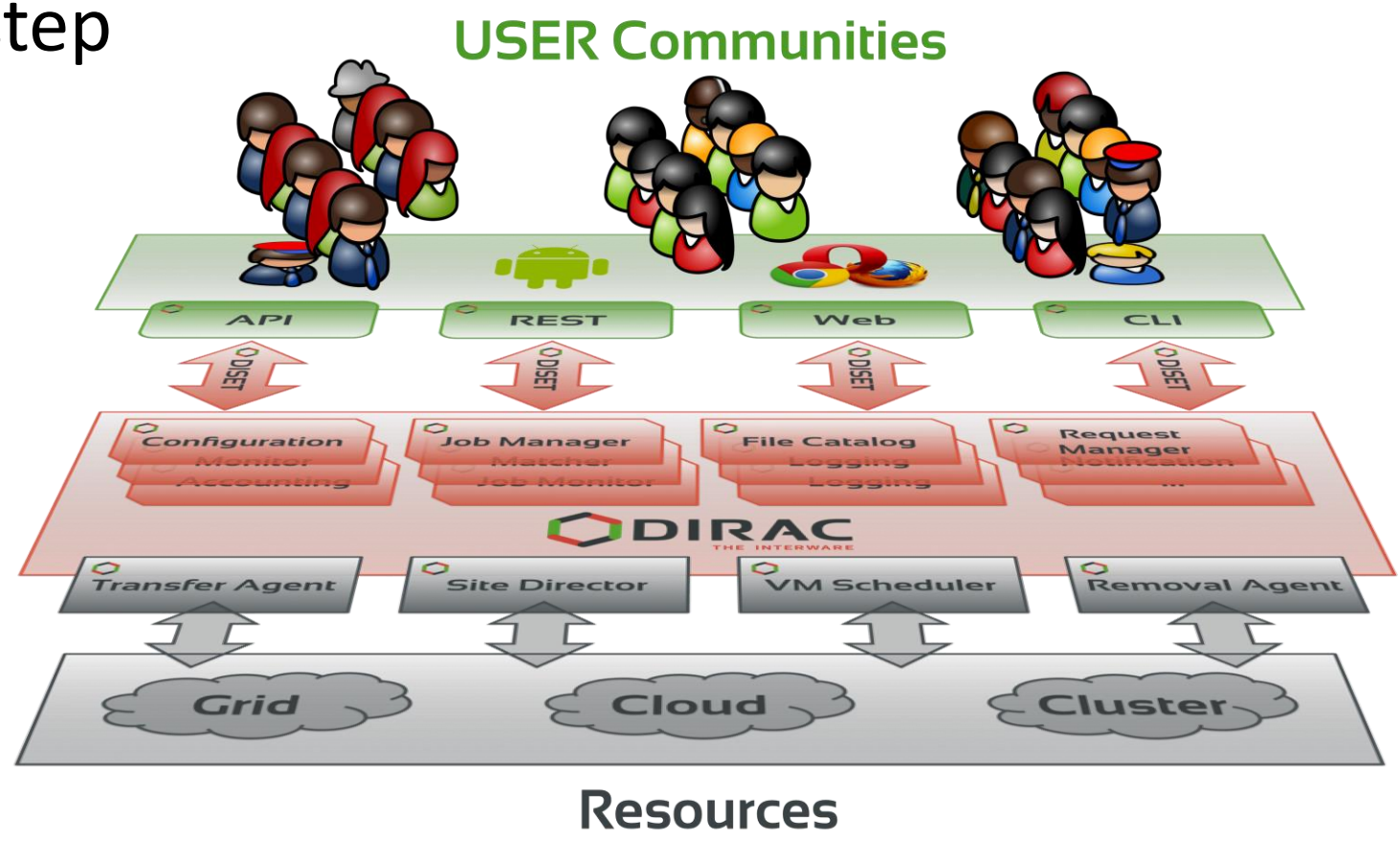
# Institut Français de Bioinformatique

- Development of an academic cloud dedicated to the management and analysis of molecular biology data
  - 10.000 cores
  - 1PO storage
- Cloud stack: Stratuslab (OpenNebula)
  - Successful prototyping at IBCP
- Testing started early 2014

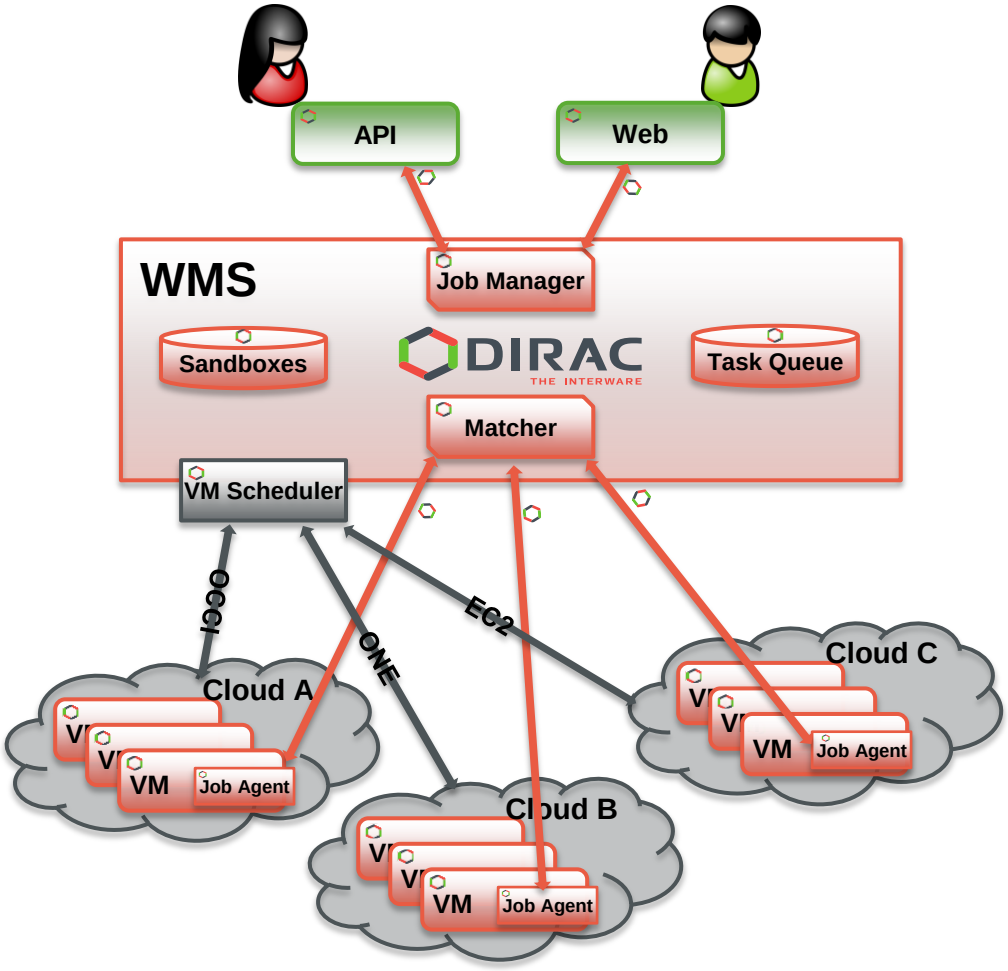


# Migration of scientific gateways from grids to clouds

- Pilot agent platforms hide the technological step



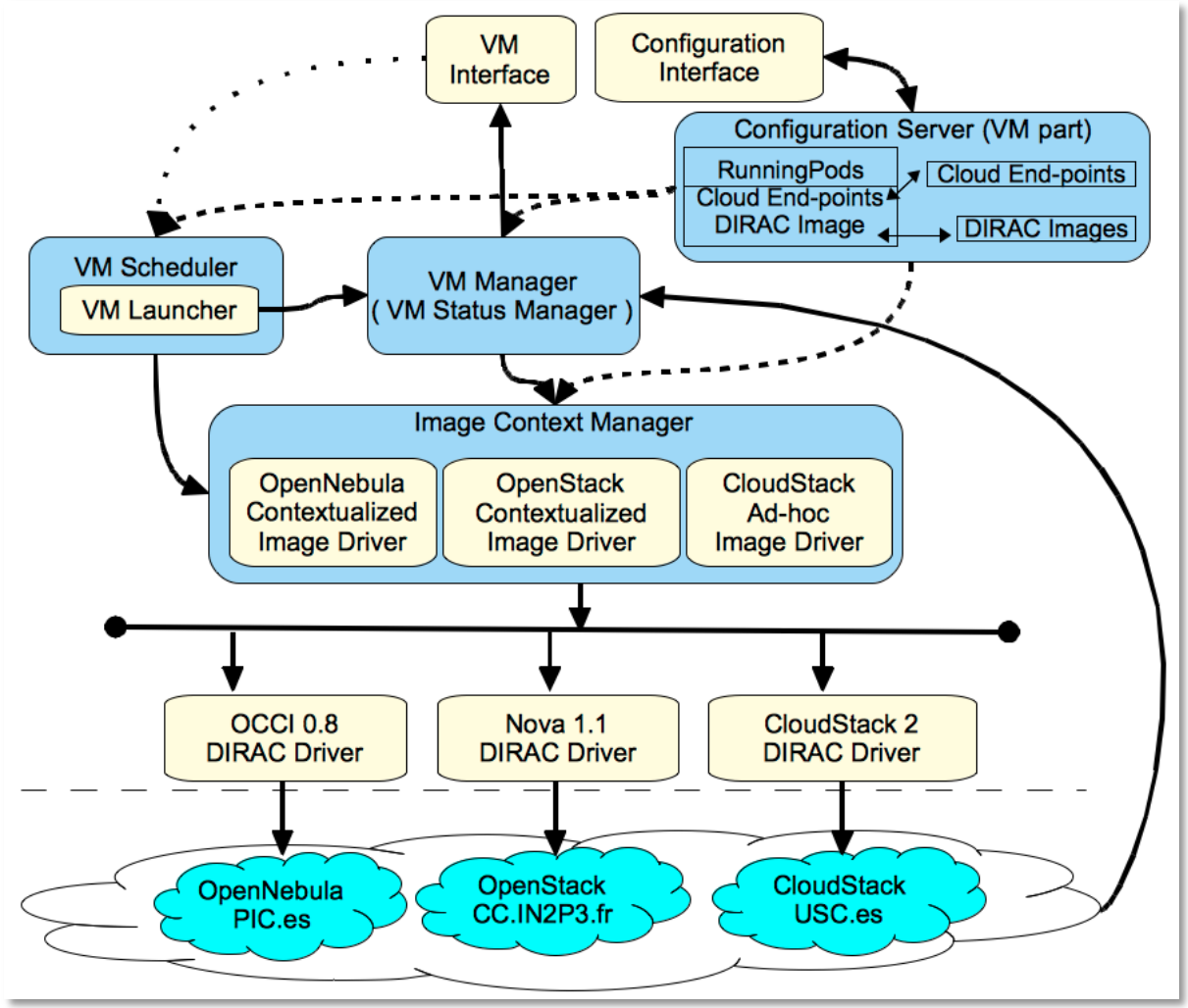
# DIRAC & Clouds






Credit: Vanessa Hamar

# Federated Cloud Test





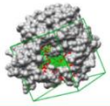

Credit: Vanessa Hamar

# WISDOM follow-up: High Throughput Computing as a Service

**HTC Problems**
**(Large amounts of computing power for lengthy period)**

Application Client 

Application Client 

Application Client 

**Web Service Interface**

**HTCaaS Server**

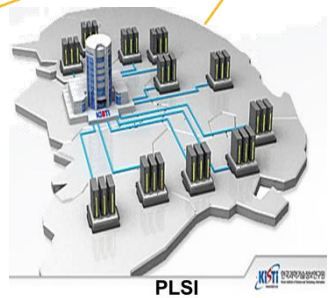
Job Manager — Job Queue

Agent Manager — Agent Submission

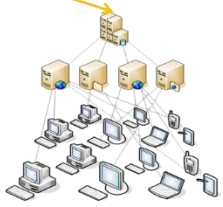
User Data Manager

**Unified Interface**


Grids


PLSI Supercomputers


Cloud

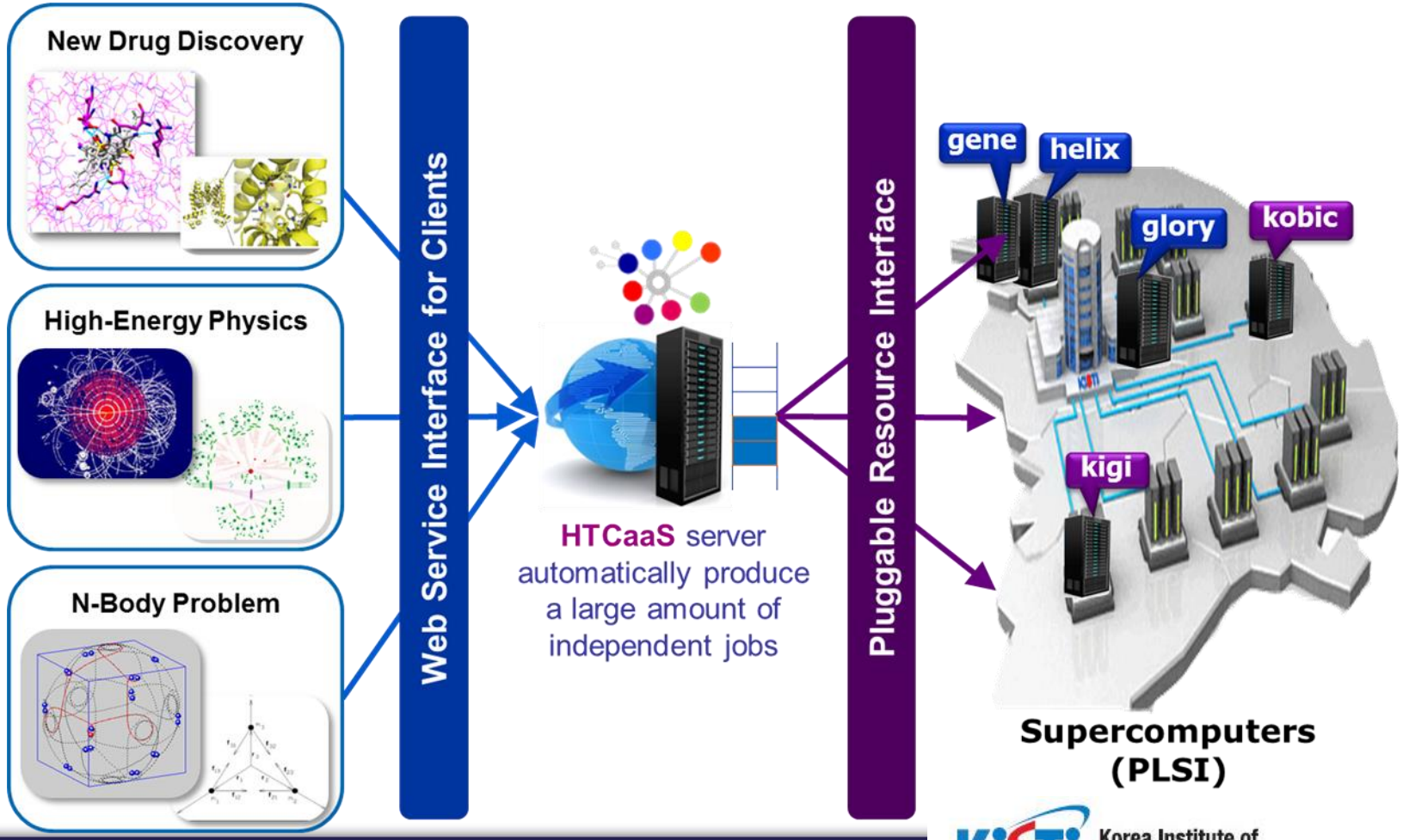

Desktop Grids

Credit: Soonwook Hwang

# PLSI: Partnership & Leadership for the nationwide Supercomputing Infrastructure

- Consortium of 14 HPC Computing Centers in Korea
- ~100 TF computing capacity by combining 17 computing resources at 9 partner sites over a dedicated high-performance network

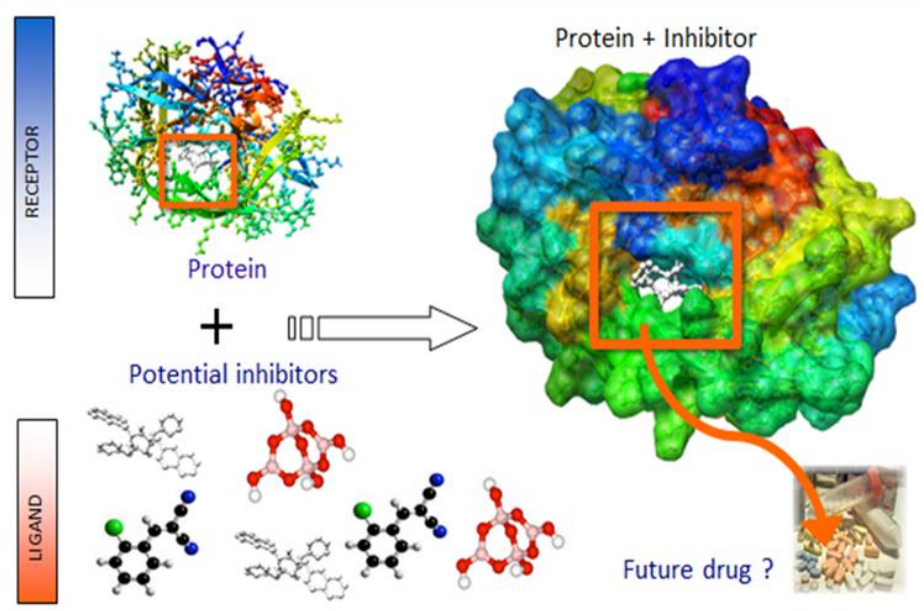
# Pilot job-based High Throughput Computing(HTC) Environment running on top of PLSI



# Protein-Ligand Docking using HTCaaS

## Virtual Screening using Molecular Docking

- Autodock3/4, a suite of automated docking tools
  - perform the docking of ligands to a set of target proteins to discover new drugs for several serious diseases such as SARS or Malaria




| No | Target Protein | PDB code | Ligand | Number of ligand | Meta Job ID | Protein preparation | |
|---|---|---|---|---|---|---|---|
| | | | | | | Gene Cloning | Protein Expression |
| 1 | Neuraminidase N1 | 3TI3 | Chembridge | 11455 | 125 | O | O |
| | | | | 39533 | 123 | | |
| | | | | 47027 | 126 | | |
| | | | | 66141 | 127 | | |
| | | | | 68880 | 128 | | |
| | | | | 75099 | 129 | | |
| | | | Natural compounds | 2720 | 124 | | |
| 2 | 3C-like protease SARS | 2ZU5 | Natural compounds | 2720 | 140 | O | O |
| 3 | Human intestinal maltase | 2QMJ | Natural compounds | 2665 | 8 | O | O |
| | | | Carbohydrate | 14473 | 29 | | |
| | | | Marine Compounds | 6154 | 25 | | |
| 4 | Malaria | 3BPF | Natural compounds | 2720 | 130 | O | O |
| | | 1YVB | Carbohydrate | 14473 | 27 | | |
| | | | Marine compounds | 6154 | 24 | | |
| | | | Natural compounds | 2665 | 6 | | |

# Key messages

- Grid computing has allowed building a truly multidisciplinary distributed IT infrastructure
- Cloud computing allows extending the grid functionalities
  - Life sciences will benefit even more
  - Public cloud prices and performances are not so appealing
  - Still a long way to the plateau of maturity for academic clouds
  - Pilot agent platforms allow a smooth transition from grids to clouds for users
    - Use of HPC resources through pilot agent platforms for High Throughput Computing