NOVEMBER 2014

CASE STUDY

Big Data HPC for Health Discovery

Key Benefits

- HPC complexity is hidden by ProActive
- Better data management
- Jobs submission from your R environment

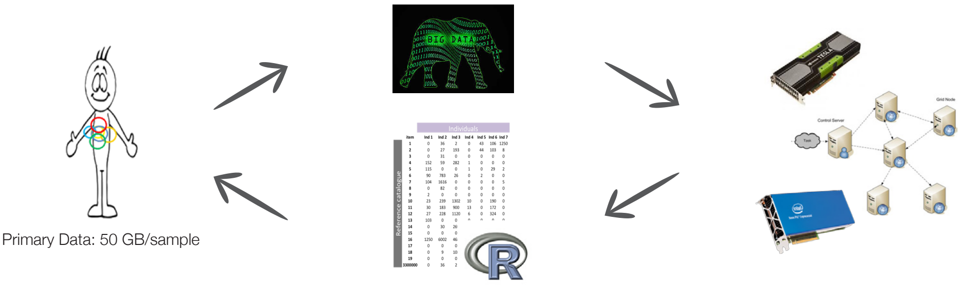
Overview

Quantitative analysis of human microbiome was developed by the INRA-MetaGenoPolis team during the EU MetaHIT (www.metahit.eu) project, which have involved a consortium of **14 institutions**. The team is now implementing a second-generation platform to support new research leveraging **'R' statistical tools with HPC** and ProActive Parallel Suite.

An innovative approach for big data processing

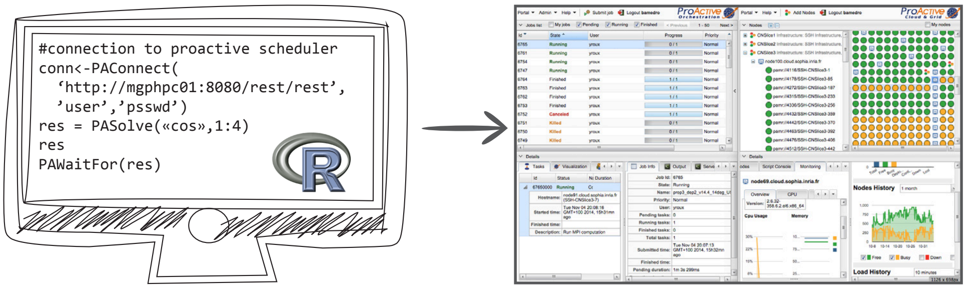
Flexible and powerful computing infrastructure as well as software are needed to analyze **big datasets** efficiently (from 40GB to 200TB). The effective use of **HPC technologies** based on new promising architectures that have emerged recently is the solution to this challenge.

Metagenopolis and ActiveEon have implemented solutions based on **Map/Reduce and jobs submissions** with 'R' ParConnector package.



ParConnector: ProActive API for R

ParConnector is a solution to the sharding with submitting jobs from R with PASolve() function of package 'R' ParConnector. Monitoring is done through submissions to the ProActive scheduler and the results are collected directly from R.



```
#connection to proactive scheduler
conn<-PAConnect(
  ‘http://mgphpc01:8080/rest/rest’,
  ’user’,’psswd’)
res = PASolve(«cos»,1:4)
res
PAWaitFor(res)
```

Submission of jobs directly from R with ProActive PASolve() function

Contacts

ActiveEon

2000, Route des Lucioles
Les Algorithmes - Pythagore B
06560 Sophia Antipolis FRANCE

Tel. +33 (0) 9 88 77 76 60
Fax +33 (0) 9 88 77 76 61

contact@activeeon.com
www.activeeon.com

INRA

MetaGenoPolis
Domaine de Vilvert - Bât 325
78352 Jouy-en-Josas FRANCE

contact@mgps.eu
www.mgps.eu

leader in human metagenomics research

Quantitatve Metagenomics focuses on the collective genome of the species composing a given ecosystem. Gut microbiota constitutes an ecosystem of major interest for the biomedical field. Quantitative Metagenomics analyses bacterial DNA diversity and helps us establish its components (bacterial species) relative abundance. It produces a measure DNA abundance on a matrix of thousands columns and ten millions rows. It's a plain matrix of floating value.

Studying so large a dataset is very challenging with conventional methods due to their high dimensionality. Bioanalysts use R, which despite being an excellent analytical and data processing platform, is not suitable for big data calculations.

Combining the excellence of GPU with the flexibility of 'R' language is a first step to support the bioanalyst using R. GpuSat is an 'R' package constructed to optimize some of the methods (Pearson, Spearman, Wilcoxon correlation) for the treatment of metagenomic data. This package aims at small scales architectures (two machines GPU with 8 devices each), as well as architectures of a larger scale using Japanese computer HAPACS/TCA