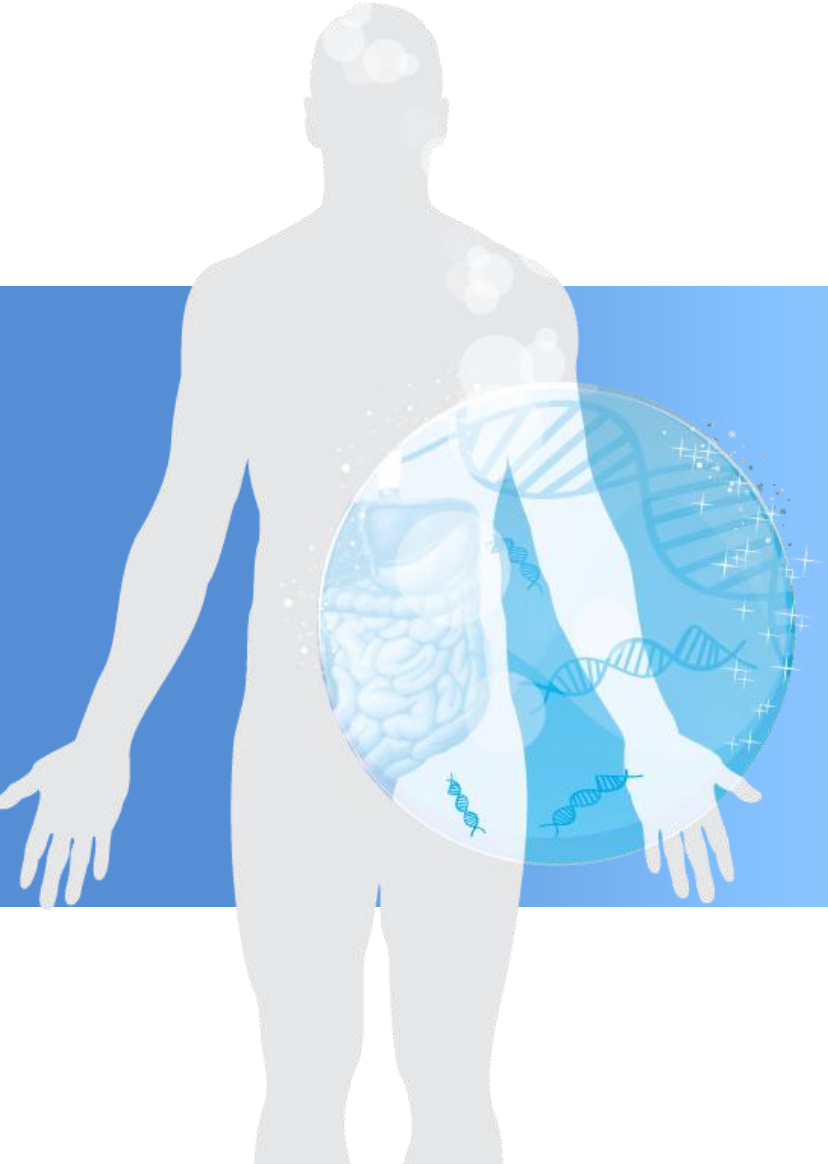
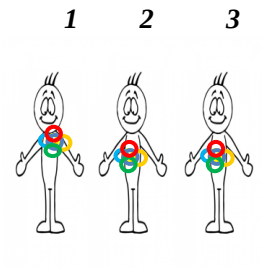


R MEETING/ JUNE 25 2014

OPTIMIZATION IN R

Ndeye Aram GAYE /Jean-Michel BATTO
Metagenopolis

MGP pipeline for metagenomic analysis

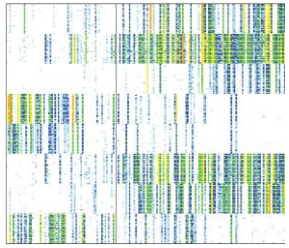


individuals

Individuals

Reference catalogue (genes)	Ind 1	Ind 2	Ind 3	Ind 4	Ind 5	Ind 6	Ind 7
1	0	36	2	0	43	106	1250
2	0	27	193	0	44	103	8
3	0	31	0	0	0	0	0
4	152	59	282	1	0	0	0
5	115	0	0	1	0	29	2
6	90	783	26	0	2	0	0
7	104	1616	0	0	0	0	5
8	0	82	0	0	0	0	0
9	2	0	0	0	0	0	0
10	23	239	1302	10	0	190	0
11	30	183	900	13	0	172	0
12	27	228	1120	6	0	324	0
13	103	0	0	0	0	0	0
14	0	30	269	0	0	0	0
15	0	0	0	0	0	95	0
16	1250	6002	468	607	492	141	8023
17	0	0	0	0	0	0	0
18	0	9	108	0	0	55	0
19	0	0	0	3	0	0	0
3300000	0	36	2	0	43	106	1250

200 ind
200000 ind



10 M
↓
40 M

BiG dAtA

BiG dAtA

Language and scientific computing environment

R code: Interpreted language

```
hello<-function(){
  .C("main.so")
}
```

C code: Compiled language

```
#include <stdio.h>
hello()
{
    printf("hello, world\n");
}
```

→ *Binary: 01001000, 01100101, 01101100, 01101100, 01101111,..*

Interpreted language VS compiled language

Compiled language : source files are converted to binary

Interpreted language : read but is not compiled.

Compiled language faster than Interpreted language

User Point of View

OpenMp → Open Multi-Processing

Message Passing Interface

Compute Unified Device Architecture

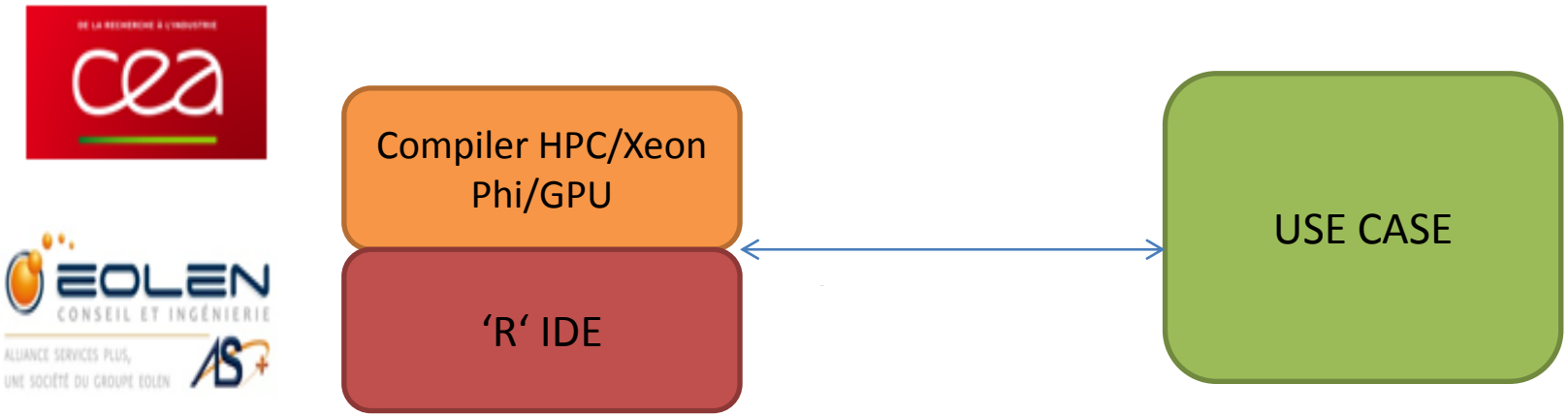
MPI

OpenCL

Cuda

Open Computing Language

✓ ***Complicated for bioanalysist***



- ITEA2 European project
- Duration 3 years

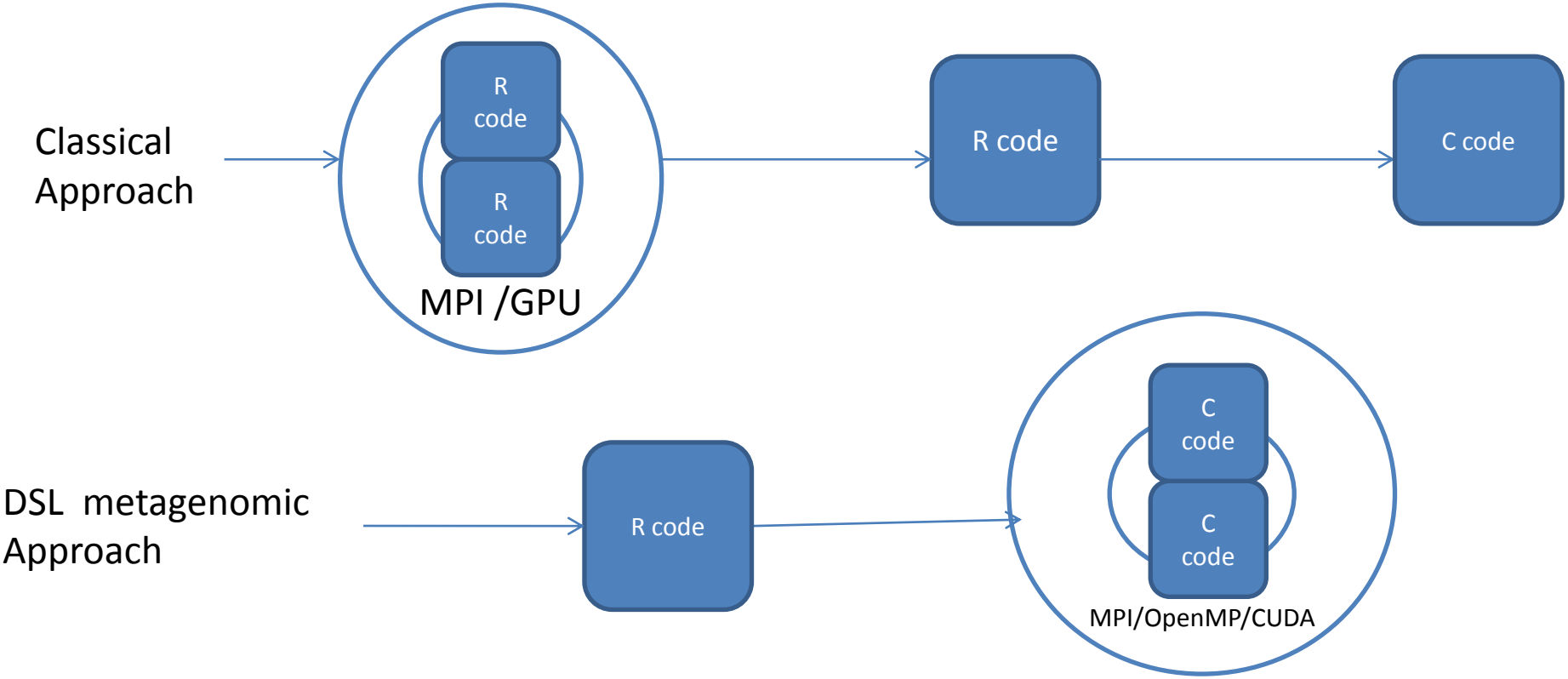
Domain Specific Language Applied to Metagenomic

- DSL: is a programming language whose specifications are dedicated to a specific application domain
- Old fashion optimization : smart library
- New fashion : DSL -> smart compiler

Strategy

Current situation:

- Need to improve calculation in R (Users)
- Acquiring experience with metagenomic dataset (Me)



MACH : New Accelerator

Xeon Phi

- ✓ 61 cores
- ✓ 244 threads
- ✓ Performance up to 1.2 teraflops

At Today's Date

✓MegaPack: Map/Reduce Package (Management of a matrix of 10M)

1 R extension (ready to deliver)

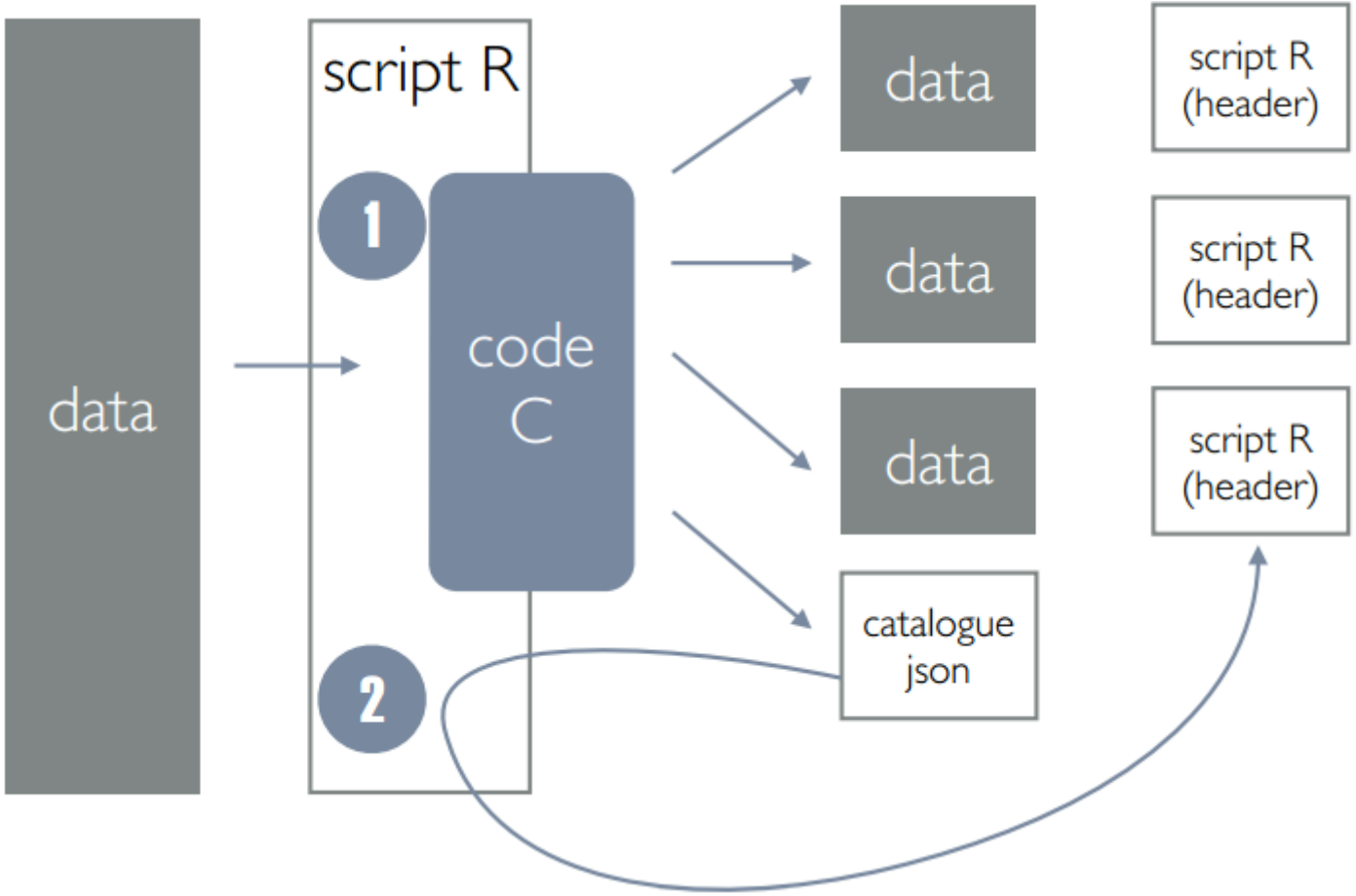
✓Parconnector: ProActive R API (Computing on Cluster)

12 R extensions (ready to deliver)

✓GpuStat: calculation of Pearson correlations in gpu (GPU utilisation)

1 R extension (ready to deliver)

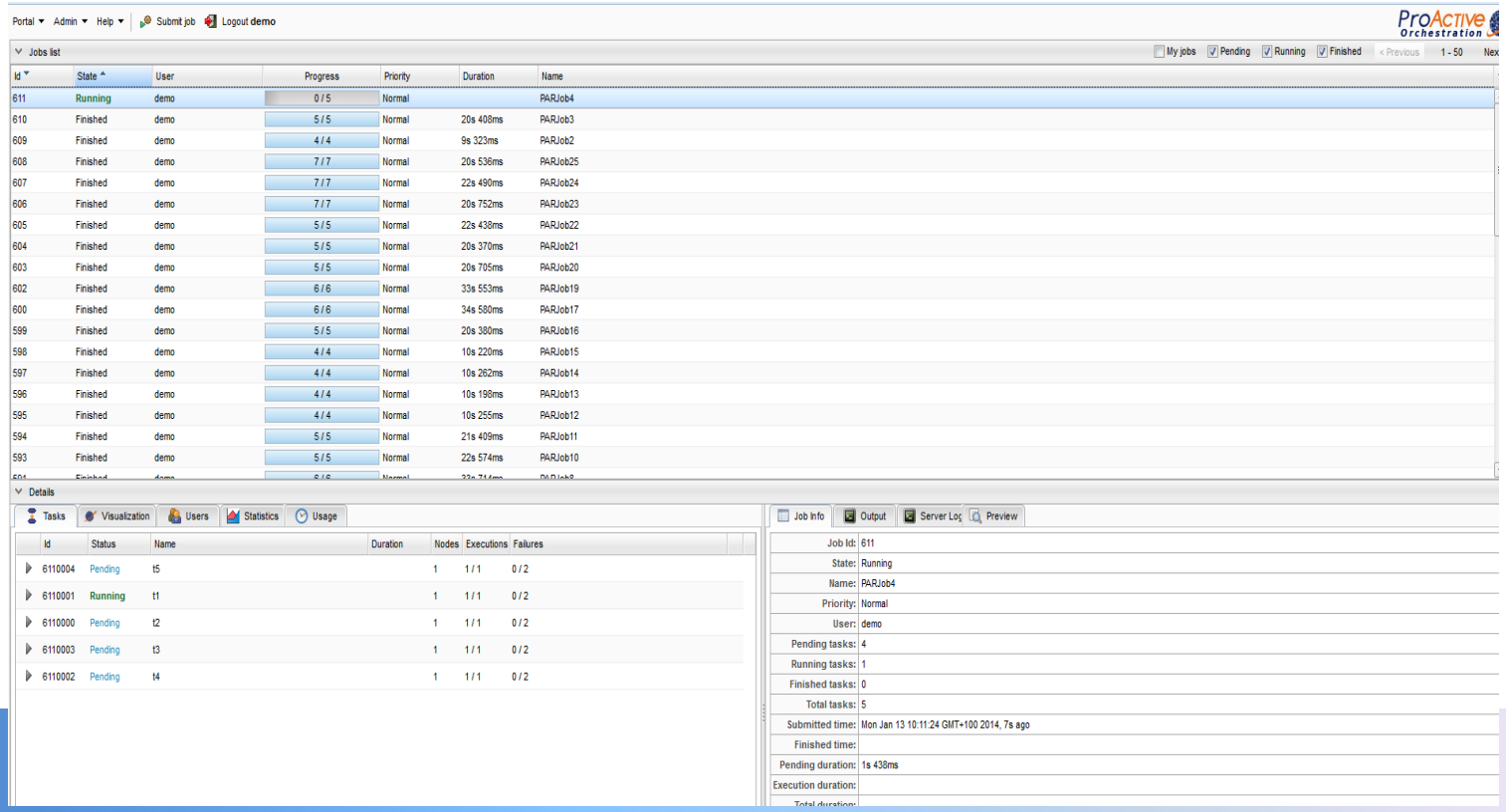
Megapack R Extension API



ProActive R Extension API

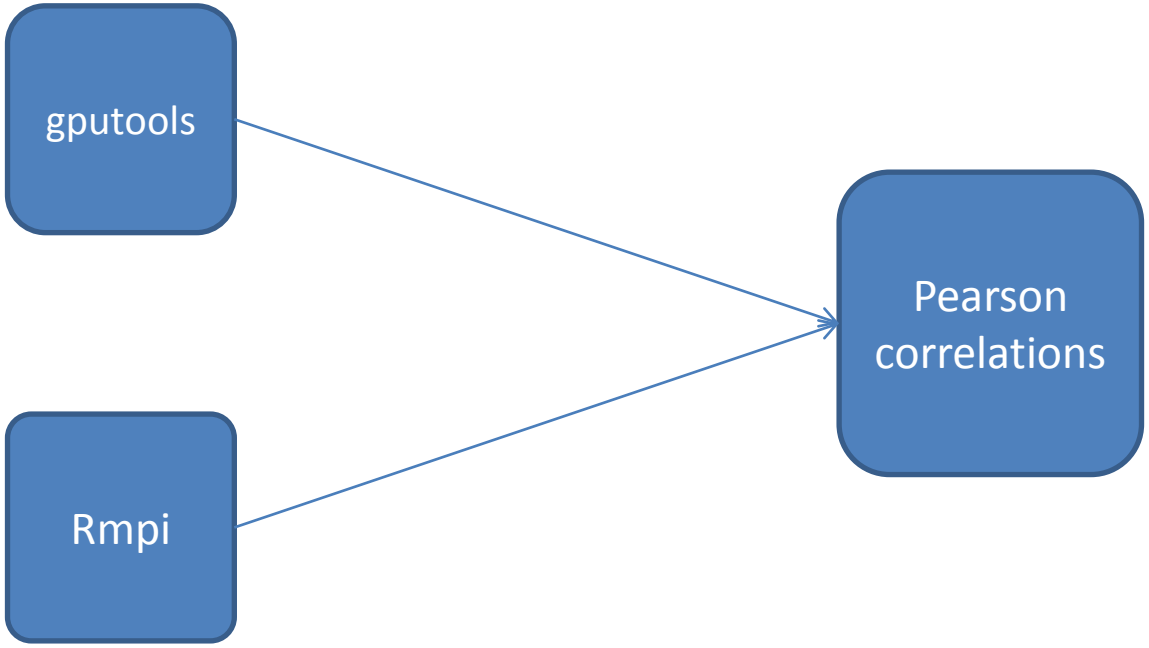
- Package PARConnector (PAConnect, PASolve, PAWaitFor, …)

  Connection with ProActive Scheduler
- To hide scheduler complexity

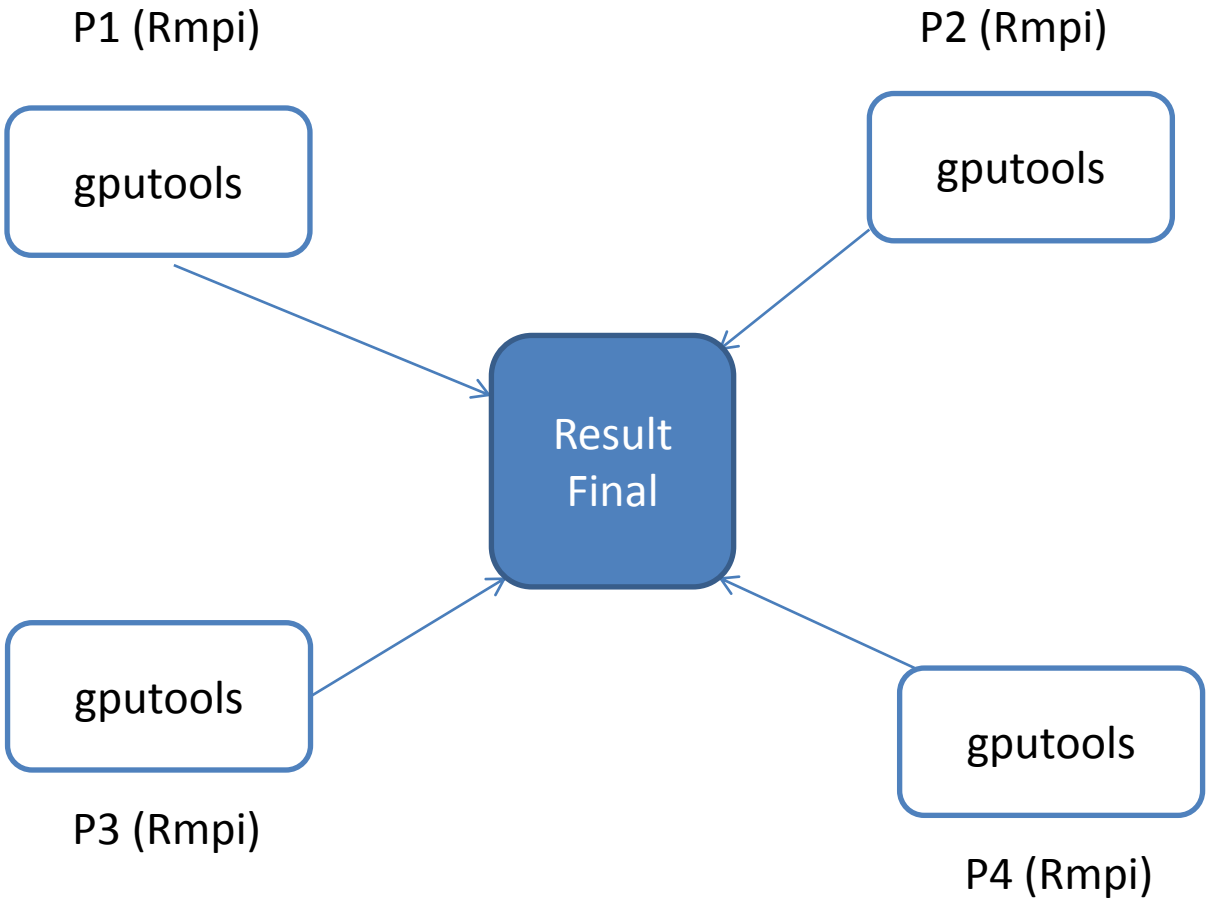


gpuStat R Extension API

Based on 2 existing R packages

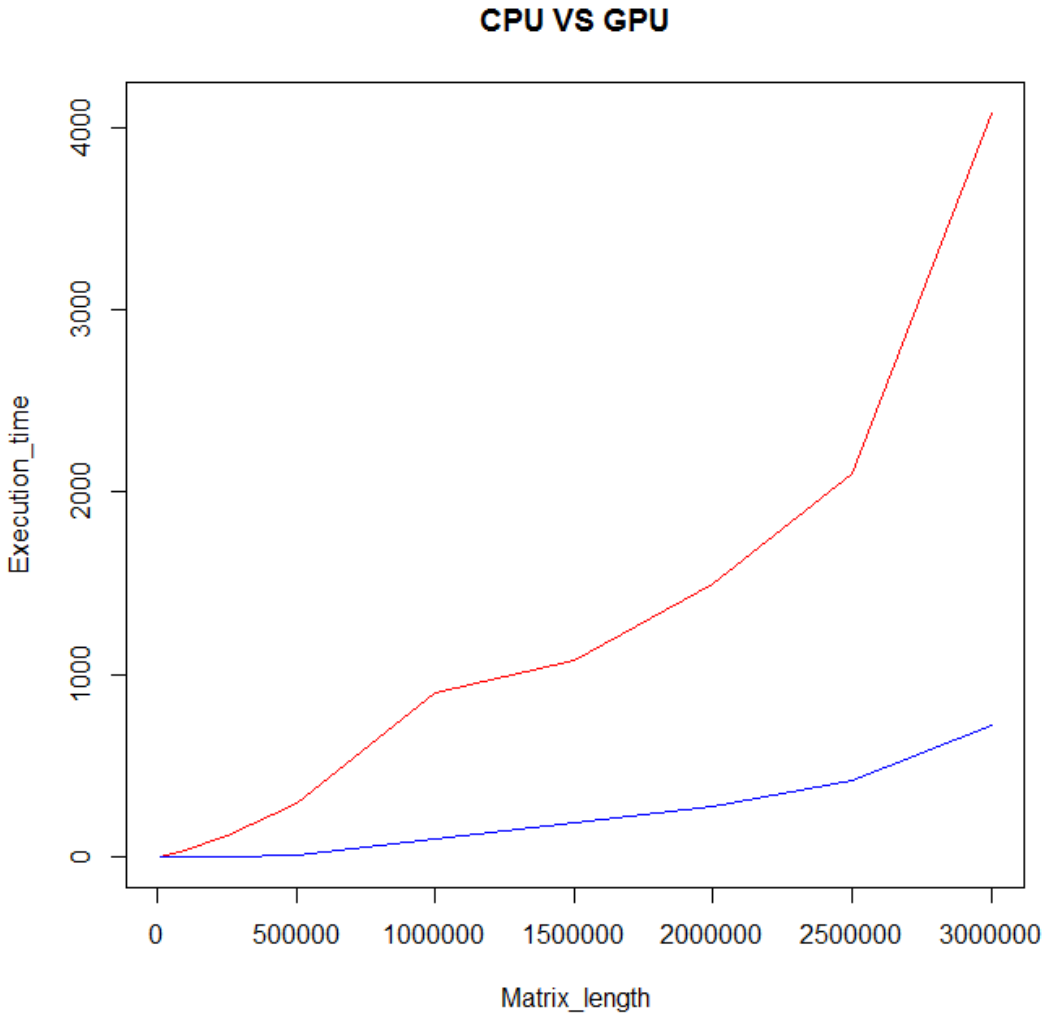


gpuStat


P1 (Rmpi)
P2 (Rmpi)
gputools
gputools
Result
Final
gputools
gputools
P3 (Rmpi)
P4 (Rmpi)

metagenopolis mgps.eu

SOME BENCHMARKS



R Use Case Applications

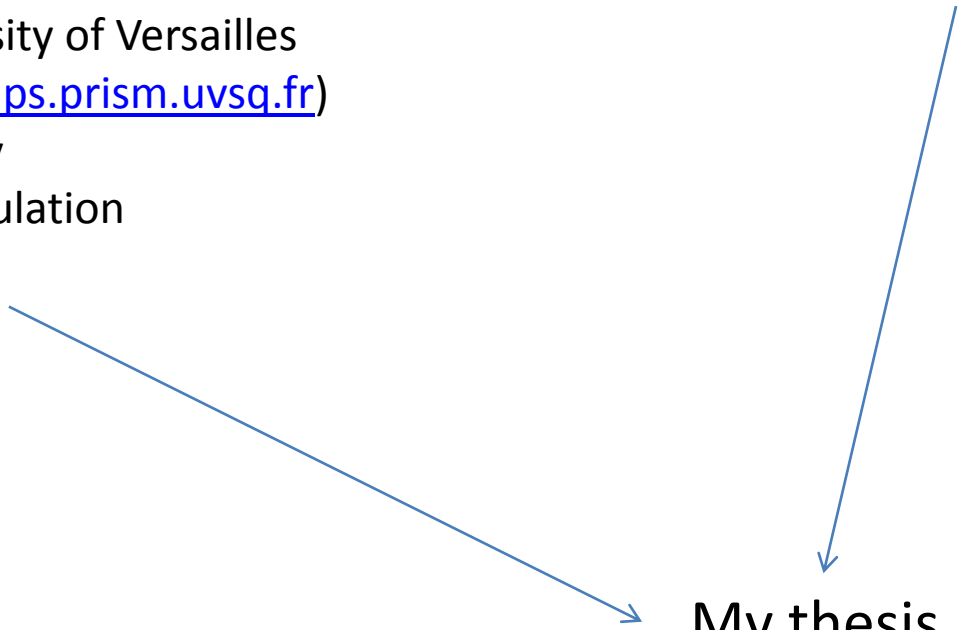
✓***MetaOmineR GPU edition (E Le Chatelier, E Prifti):***
data analysis of quantitative metagenomics

✓***Bayesian Builder (J Abou Ghantous, J Tap):*** bayesian estimates optimization applied to data from the study of the human intestinal microbiota.

Beyong MACH (HPC)

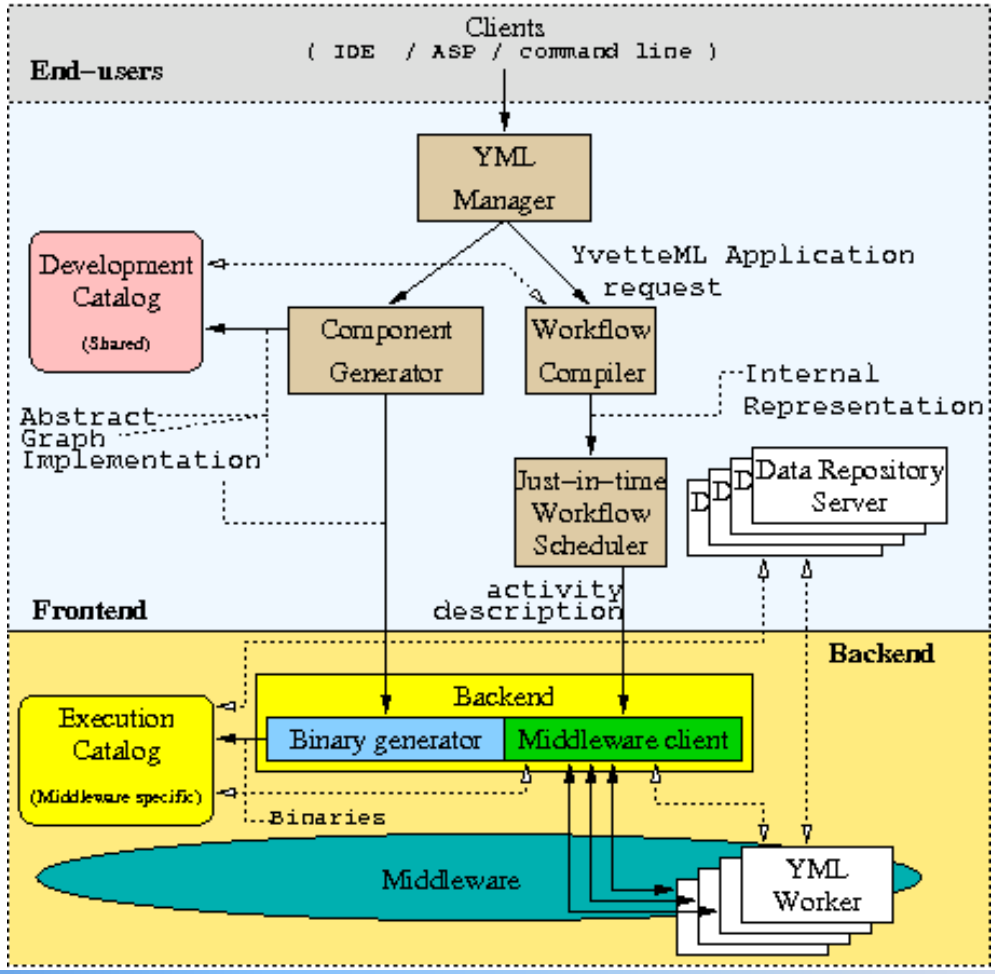
Nahid Emad
Professor, University of Versailles
Resp. MIHPS (mihps.prism.uvsq.fr)
PRiSM Laboratory
Maison de la Simulation

Datacenter: YML
(YvetteLanguage)



My thesis

YML (yml.prism.uvsq.fr)



metagenopolis mgps.eu

Conclusion

✓Our strategy allows us to start quicky

✓ Next milestone R Xeon Phi package

✓ Next milestone R in YML

Thanks

Jad Abou Ghantous
Anne-Sophie Alvarez
Jean-Michel Batto
Amine Ghozlane
Vincent Heuschling
Emmanuelle Le Chatelier
Pierre Léonard
Nicolas Pons
Edi Prifti
Julien Tap