R MEETING/ JUNE 25 2014
OPTIMIZATION IN R

Ndye Aram GAYE /Jean-Michel BATTO
Metagenopolis
MGP pipeline for metagenomic analysis

10 M

40 M

200 ind

200000 ind

BiG dAtA

BiG dAtA
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

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

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

Binary:
```
01001000, 01100101, 01101100, 01101111, ...
```
User Point of View

Message Passing Interface

MPI

Open Mp

Open Multi-Processing

OpenCL

Compute Unified Device Architecture

Cuda

Open Computing Language

✓ Complicated for bioanalyst
- 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
Current situation:
- Need to improve calculation in R (Users)
- Acquiring experience with metagenomic dataset (Me)
MACH : New Accelerator

- 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

- gputools
- Rmpi

Pearson correlations
**gpuStat**

- P1 (Rmpi)
  - gputools

- P2 (Rmpi)
  - gputools

- P3 (Rmpi)
  - gputools

- P4 (Rmpi)
  - gputools

**Result Final**
SOME BENCHMARKS

CPU VS GPU

Execution time

Matrix length
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)
Conclusion

✓ Our strategy allows us to start quickly

✓ 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