Parallel Computing in R using the BoT Package

Florent Chuffart

Laboratoire de Biologie Moléculaire de la Cellule
Ecole Normale Supérieure de Lyon
UMR5239 CNRS/ENS Lyon/UCBL/HCL
46, allée d'Italie
69364 Lyon cedex 07
florent.chuffart@ens-lyon.fr

Mots clefs : Distributed Computing, Bag-of-Tasks, Parameter Sweep.

BoT (stands for Bag Of Tasks) is an R package allowing to distribute independent tasks over many cores and many computing nodes. The simple fact that BoT is based on the process forking feature and task locking over file system makes BoT compatible with most of computing infrastructures: multicore, clusters, grids and clouds (see Figure 1).

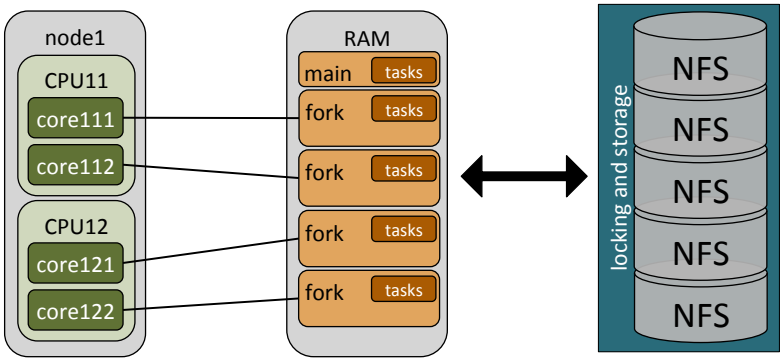

Figure 1: BoT architecture.

Using BoT, each task is a set of parameters associated with a user-defined function built on an R process. Next step consists in forking this R process for each core of the computing node. Finally, the forked set of tasks is randomized and executed in a parallel way. When a task starts a distributed lock is taken. This avoids redundant task execution. When a task is ended, result is dumped into a file.

As R package *mapReduce*, BoT uses a flexible parallelization backend. On the other hand BoT it isn't restricted to the MapReduce computational paradigm. Unlike R package *multicore* that is based on shared memory paradigm, BoT manages distributed memory: BoT is designed to be run on many heterogeneous computing ressources. BoT is based on the R package *fork*, it extends it in a fair way.

BoT is used to analyse ChIP-Seq data in the SiGHT project context (ERC-StG2011-281359). BoT has been used on two computing infrastructures: Grid'5000 experimental testbed [1] and PSMN computing center of ENS de Lyon [2]. BoT package is available on our web page[3].

[1] `https://www.grid5000.fr`
[2] `http://www.ens-lyon.fr/PSMN`
[3] `http://www.ens-lyon.fr/LBMC/gisv/index.php/en/protocols/bioinformatics`