


R em ambiente HPC

Escola Supercomputador Santos Dumont 2018
Laboratório Nacional de Computação Científica – LNCC
Instrutores: Raquel Lopes Costa & Guilherme Gall

Página do curso:  https://github.com/quelopes/R_for_HPC

Resumo

O R é uma linguagem bastante utilizada por cientistas de dados, estatísticos, biólogos, epidemiologistas e funciona muito bem em muitas pesquisas, gerando resultados de qualidade em relatórios e artigos científicos. A flexibilidade e facilidade de programar avançando nas análises a partir de funções e pacotes disponíveis são vantagens decisivas na escolha dessa linguagem por pesquisadores.

No entanto, a análise de dados massivos decorrentes principalmente da era 'Big Data' envolvem avançados algoritmos com grande número de parâmetros necessários para ser estimado. Analisar os dados massivos tem sido um fator limitante na linguagem R. Isso porque o R foi originalmente pensado para ser executado em uma única *thread*, além de ser uma linguagem interpretada, sendo a maior parte das instruções executadas diretamente. Aliado a esses fatores, o conhecimento sobre MPI, SOCKET, C, C++ e Fortran são barreiras para grande de pesquisadores que lidam com cálculos estatísticos intensivos e estão tentando acelerar os cálculos implementando algoritmos paralelos.

Ao longo dos anos, diversas iniciativas tem sido feitas para melhorar a eficiência e desempenho das aplicações em R. Esse curso tem como proposta apresentar algumas dessas estratégias de paralelismo e melhora na eficiência para ganho de escalabilidade na linguagem R.

Pré-requisitos

- R básico
- Conhecimentos em plataformas HPC

Ementa

- R, introdução geral
 - O que é o R
 - Vantagens do R
- Funções da família apply
- Pacotes de paralelismo do R
 - Pacote parallel (funções análogas às da família apply)
 - Pacote foreach e backends paralelos (estrutura análoga ao laço for nativo do R)
- Benchmark no R
- Exemplos com análise de agrupamento
- Otimização do código
- Estudo de caso, Model-R

Bibliografia

- State of art in parallel computing with R. Journal of Statistical Software. August, 2009. Vol. 31(1).
- A survey of R software for parelle computing (2014) Vol. 2(4). Americal Journal of Applied Mathematics and Statistics.
- Ethan McCallum e Stephen Weston (2012). Parallel R. (Book). O'Reilly.
- Lin, H., Yang, S. and Midkiff, S. P. (2013). A Parallel R Framework for Processing Large Dataset on Distributed Systems. Dataset on Distributed Systems, DataCloud.