High Performance Computing for R using SPRINT: from multi-cores to the cloud

Muriel Mewissen[1,*], Thorsten Forster[1], Terence Sloan[2], Michal Piotrowski[2], Lawrence Mitchell[2], Peter Ghazal[1], Arthur Trew[2], Jon Hill[3]

1. Division of Pathway Medicine, The University of Edinburgh, Chancellor's Building, 49 Little France Crescent, Edinburgh, EH16 4SB, UK.
2. Edinburgh Parallel Computing Centre, James Clerk Maxwell Building, The King's Buildings, Mayfield Road, Edinburgh, EH9 3JZ, UK.
3. AMCG, Earth Science and Engineering, Imperial College, London, SW7 2AZ, UK.

*Contact author: Muriel.Mewissen@ed.ac.uk

Keywords: High Performance Computing, SPRINT, bioinformatics, bioconductor, Cloud

The SPRINT (Simple Parallel R INTerface) package allows R users to exploit High Performance Computing (HPC) easily. It is particularly targeted at biostatistians processing the ever increasing volumes of post-genomic data. SPRINT requires minimal HPC knowledge and minimal changes to existing R scripts.

HPC is available in many forms from multi-core PCs to supercomputers. SPRINT is implemented using the MPI [1] standard, a parallel processing technology supported by a wide range of platforms. This ensures SPRINT can run and provide performance benefits on multi-core desktop PCs, shared memory platforms, clusters, clouds and supercomputers [2].

Previous work has demonstrated extremely good performance and scalability for the SPRINT implementations of the R clustering (**pam**), permutation testing (**mt.maxT**) and Pearson correlation (**cor**) functions. In particular, the SPRINT implementation of permutation testing has close to optimal scaling on up to 512 processors on a supercomputer [3]. To address further analysis bottlenecks highlighted by the R user community [4], recent additions to the HPC functionality provided in SPRINT have included versions of the standard R **apply** and **boot** functions, the machine learning **random forest** function, and the general statistical analyses **rank product** function.

This paper will present our investigation into the performance of the SPRINT parallel R functions on various HPC architectures including clouds.

References

[1] L. Clarke et all (1994). The MPI Message Passing Interface Standard. Programming environments for massively parallel distributed systems: working conference of the IFIP WG10.3, Ed. K.M.Decker and R.M.Denham, Birkhauser 1994, ISBN 3-7643-5090-3.

[2] B. Dobrzelecki et all (2011). Managing and Analyzing Genomic Data using HPC and Clouds, book chapter in G. Aloisio & S. Fiore "Grid and Cloud Database Management", Springer.

[3] S. Petrou et all (2010). Optimization of a parallel permutation function for the SPRINT R Package. HPDC2010 Proceedings, http://salsahpc.indiana.edu/ECMLS2010/papers/064.pdf.

[4] M. Mewissen (2009). SPRINT – User Requirement Survey Results, http://sprint.gti.ed.ac.uk/Docs/EXTERNAL_SPRINT-URSR_v1.1.pdf.