Elevating R to Supercomputers

Drew Schmidt[1,*], **Wei-Chen Chen**[2], **Pragneshkumar Patel**[1], **George Ostrouchov**[1,2]

1. Remote Data Analysis and Visualization Center — University of Tennessee Knoxville, USA
2. Computer Science and Mathematics Division — Oak Ridge National Laboratory, USA
[*]Contact author: schmidt@math.utk.edu

Keywords: parallel computing, high performance computing, big data, SPMD

The biggest supercomputing platforms in the world are distributed memory machines, but the overwhelming majority of the development for parallel R infrastructure has been devoted to small shared memory machines. Additionally, most of this development focuses on task parallelism, rather than data parallelism. But as big data analytics becomes ever more attractive to both users and developers, it becomes increasingly necessary for R to add distributed computing infrastructure to support this kind of big data analytics which utilize large distributed resources.

The *Programming with Big Data in R* (pbdR) project aims to provide such infrastructure, elevating the R language to these massive-scale computing platforms. The main goal of the project is to empower data scientists by bringing flexibility and a big analytics toolbox to big data challenges, with an emphasis on productivity, portability, and performance. We achieve this in part by mapping high-level programming syntax to portable, high-performance, scalable, parallel libraries such as MPI and ScaLAPACK. This not only benefits the R community by enabling analysis of larger data than ever before with R, but it also benefits the supercomputing community which, to date, mostly only dabbles superficially with statistical techniques.

A major focus of the project is ease of use, with great effort spent towards minimizing the burdens of supercomputing for R users and developers. Programs written using pbdR packages are written in the Single Program/Multiple Data, or SPMD style (not to be confused with SIMD architecture computers), which is a very natural extension of serial programming for distributed platforms. This paradigm together with extensive use of R's S4 methods allows us to create highly scalable tools with nearly-native serial R syntax.

In this talk, we will discuss some of the early successes of the pbdR project, benchmarks, challenges, and future plans.

References

Blackford, L. S., J. Choi, A. Cleary, E. D'Azevedo, J. Demmel, I. Dhillon, J. Dongarra, S. Hammarling, G. Henry, A. Petitet, K. Stanley, D. Walker, and R. C. Whaley (1997). *ScaLAPACK Users' Guide*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Chen, W.-C., G. Ostrouchov, D. Schmidt, P. Patel, and H. Yu (2012). pbdMPI: Programming with big data – interface to MPI. R Package, URL http://cran.r-project.org/package=pbdMPI.

Gropp, W., E. Lusk, and A. Skjellum (1994). *Using MPI: Portable Parallel Programming with the Message-Passing Interface*. Cambridge, MA, USA: MIT Press Scientific And Engineering Computation Series.

Ostrouchov, G., W.-C. Chen, D. Schmidt, and P. Patel (2012). Programming with big data in R. URL http://r-pbd.org/.

Schmidt, D., W.-C. Chen, G. Ostrouchov, and P. Patel (2012). pbdDMAT: Programming with big data – distributed matrix algebra computation. R Package, URL http://cran.r-project.org/package=pbdDMAT.