Running R with 120 threads on the Intel® Xeon® E7-4870 v2

Eric Kramer[1*], William Shipman[1], Ali Torkamani[1]

1. Scripps Translation Science Institute, 3344 North Torrey Pines Court, Suite 300, La Jolla, CA 92037
*Contact author: ekramer@scripps.edu

Keywords: Parallel Processing, Machine Learning, Big Data, High Performance Computing

The increasing size of datasets and the proliferation of multicore CPUs has increased demand for parallel processing in *R*. Existing packages, such as the **parallel**, **snow** [6]**, foreach** [4] and **multicore** [8] packages, allow users to parallelize loops in *R*. Similarly, several groups have demonstrated impressive performance gains by compiling *R* with multithreaded BLAS libraries, such as Intel's Math Kernel Library [5]. We use these strategies to deploy *R* on a 60-core server, which is capable of operating 120 concurrent threads on four Intel Xeon E7-4870 v2 CPUs. Using Urbanek's benchmarking script [7], we see a 230-fold increase in performance for calculating cross products as compared to the base *R* installation, but only a 1.7-fold performance increase for sorting random numbers. We also trained tumor classifiers using methods from the **nnet** [9], **kernlab** [2], and **caret** [3] packages with data from The Cancer Genome Atlas [1]. We see a 20-fold performance gain for training support vector machines as compared to the base R installation, and a 21-fold performance gain for training neural networks with three layers. Overall, these findings suggest that using dozens of threads can result in large performance gains for matrix operations, support vector machines and neural networks in *R*.

References

[1] The Cancer Genome Atlas Network (2012). Comprehensive molecular protraits of human breast tumors. Nature 490(7418), 61-70.

[2] Alexandros Karatzoglou, Alex Smola, Kurt Hornik, Achim Zeileis (2004). kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software 11(9), 1-20. http://www.jstatsoft.org/v11/i09/

[3] Max Kuhn (2014). Caret: Classification and Regression training. http://cran.r-project.org/web/packages/caret/index.html

[4] Revolution Analytics (2013). Foreach: foreach looping construct for R. http://cran.r-project.org/web/packages/foreach/index.html

[5] Revolution Analytics (2013). High performance R. http://www.revolutionanalytics.com/high-performance-r

[6] Luke Tierney, A. J. Rossini, Na Li, H. Sevcikova (2013). Snow: Simple Network of Workstations. http://cran.r-project.org/web/packages/snow/index.html

[7] Simon Urbanek (2008). R Benchmarking Script. http://r.research.att.com/benchmarks/R-benchmark-25.R

[8] Simon Urbanek (2011). Multicore: Parallel processing of R code on machines with multiple cores or CPUs. http://cran.r-project.org/web/packages/multicore/index.html

[9] W. N. Venables, B. D. Ripley (2002). Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0