BOF Report: Super-R: Supercomputing and R for Data-Intensive Analysis

R (http://www.r-project.org/) is a free, powerful, open source software package with extensive statistical computing and graphics capabilities. R has become popular for data analysis in many fields, drawing power from its high-level expressiveness and numerous domain-specific packages. While R is clearly a "high productivity" language, it is not known as a "high performance" language. However, recent efforts have resulted in methods for effectively scaling R to the power of supercomputers. The ultimate goals of this BOF are: 1) to help build a community of users, developers, and experts interested in applying R to solve data intensive problems, and 2) to explore areas where potential improvements will benefit the community's data analysis needs.

The BOF was held on Nov. 20 from 5:30 to 7pm with 43 attendees. The first half of the BOF session included four short presentations by Eric Wernert from Indiana University's Pervasive Technology Institute (PTI), Niall Gaffney from Texas Advanced Computing Center (TACC), George Ostrouchov from both the Oak Ridge Leadership Computing Facility (OLCF) and University of Tennessee's National Institute for Computational Sciences (NICS), and Weijia Xu from TACC. The four presentations covered an introduction to the high performance computing environment available through XSEDE, to the current landscape of parallel packages in R and the recent pbdR developments, and to experiences in using R with Xeon Phi processors on the Stampede cluster at TACC. The audience was very interested in the presentations, several questions regarding resources availability, code development, and optimization were raised and discussed following the presentations.

The second half of the BOF session was an interactive panel session with the audience. There were five panelists Hui Zhang, Huian Li, GuangChen Ruan (all from PTI), Drew Schmidt (NICS), and Matt Vaughn (TACC). Panelists spent several minutes sharing their experiences of using R with HPC for data intensive scientific workflows ranging from biology, to digital humanities and finance. The second part of BOF was particularly focused on three questions: *advantages to using R with HPC; the biggest difficulty in using R with HPC; and major challenges to improving the scalability of R.* More than 10 people from the audience shared their experiences and took part in the discussions. In general consensus, participants think the advantages of R come from its domain-specific design for data analysis, the availability of a wide variety of domain packages, and its comprehensive features beyond statistical analysis that include sophisticated graphics. The difficulties of using R include uneven documentation support, compatibilities and inconsistency between packages, lack of tutorials and practical guidance in developing and using R code. The scalability of R can be improved from several aspects, such as improving support from CRAN for improving the R's core-packages, providing user guidance and education, optimizing and publicizing existing HPC resources for R use.

The BOF session confirmed a general interest in bringing R and supercomputing together. The interests come across the R user community, the R developer community, and the HPC community. As we recognize prevailing needs in the three well-established communities, we also think there is a need in developing a community specifically for using R with HPC. This community can share information, experiences and collaboration in improving R use with HPC from multiple aspects. As a result of the BOF, we are starting a Google group super-r@googlegroups.com and inviting interested participants and interested readers to subscribe. We are also planning future events for the theme in the next several months.