The State of High Performance Computing in the Open Source R Ecosystem

Drew Schmidt

R is a strange language. Dating back to S from Bell Labs, it is the mad science experiment produced by blending a C-inspired programming language with a feature-rich, interactive data analysis package. It has been primarily developed by statisticians, and has an eclectic mix of programming idioms and syntax styles. One writer describes R as "the most shockingly dreadful and most useful language for data analysis".

Yet in spite of (or as the above quote suggests, perhaps because of) its many quirks, it is beloved by many data scientists. Indeed, it is the de facto standard for data analysis in academia, and has been steadily gaining popularity in industry for some time. Recently, the IEEE Spectrum programming language rankings listed R as the fifth most popular programming language. A humble scripting language designed only to be good at data analysis beat out standards like C# and Javascript in a general puprose "language shootout".

So it would seem that R is here to stay. And that includes on the cluster! Unsurprisingly, in the age of big(ger) data, statisticians, scientists, and all other analyzers of data are increasingly finding themselves in the need of HPC resources. And when they need to move to small campus clusters, national supercomputing resources, or the cloud, they want to bring R with them. But R was built with the desktop, not the cluster, in mind.

To address this, the open source R community has steadily been developing solutions to transform R from merely being a "high productivity" language, into a legitimate high performance language. These external packages enhance R computations to use multi-threaded compiled kernels, access coprocessor cards like gpu's and the Intel Xeon Phi, and even elevate R to large distributed resources, living atop technologies like MPI and Spark.

In this talk, we will explore this package landscape, and attempt to describe both the history of R's usage on HPC resources, as well as the current state of the art. We hope to provide for the attendee the proper context and motivation to understand where R currently stands in the world of HPC, and to hopefully be able to better guess at where it is going.

Speaker Bio

Drew Schmidt is a developer for the pbdR project for distributed computing with R. Currently a graduate student at the University of Tennessee, he holds an M.Sc. in mathematics from the University of Tennessee, and was formerly a research associate at the National Institute for Computational Sciences.