Programming with Big Data in R

Drew Schmidt[1,⋆], **George Ostrouchov**[2,1,†]

1. Joint Institute for Computational Sciences, University of Tennessee
2. Oak Ridge National Laboratory
⋆schmidt@math.utk.edu, †ostrouchovg@ornl.gov

Overview: The tutorial will introduce attendees to high performance computing concepts for dealing with big data using R, particularly on large distributed platforms. We will describe the use of the "programming with big data in R" (**pbdR**) package ecosystem by presenting several examples of varying complexity. Our packages provide infrastructure to use and develop advanced parallel R scripts that scale to *tens of thousands* of cores on supercomputers but also provide simple parallel solutions for multicore laptops.

Our packages are described in a textbook-style vignette associated with our package pbdDEMO. This tutorial will follow many of the examples presented in the document, which we continue to update. We conducted this tutorial successfully at UseR 2013. However, this year's tutorial will include numerous updates and improvements, including demonstrations of some of our new developments and applications completed since last year.

In this tutorial, we will:

- Provide a quick overview of parallel R's capabilities, and discuss how R can interface with parallel hardware and HPC libraries.
- Discuss the value of profiling, and show off our new profiling package pbdPAPI.
- Introduce basic MPI programming concepts, and its simplified interface via pbdMPI.
- Discuss parallel data input.
- Introduce distributed matrices and distributed matrix methods.

Attendee background: We assume intermediate knowledge of R. No prior parallel programming experience is necessary. If you wish to follow along on your multicore laptop during the tutorial, please install (or check that you have):

- R (and Rtools if you are a Windows user)
- An MPI library
- the **pbdR** packages

See our installation instructions for details about how to install these requisites for each major platform. Please note that we are anticipating having new releases of most of our packages just before UseR! 2014, so you may wish to wait until just before the tutorial to install these packages.

Workshop Materials: Slides and source code for the tutorial will be made available by June 30, 2014 on the **pbdR** website.

Presenters: Drew Schmidt is a researcher at the University of Tennessee interested in the intersection of mathematics, statistics, and high performance computing, and is one of the lead developers of the **pbdR** project.

George Ostrouchov is Senior Research Scientist at the Oak Ridge National Laboratory and Joint Faculty Professor at the University of Tennessee. George's interests are focused on the interaction of high performance computing and statistics, and he is the architect and lead of the **pbdR** project.