User friendly distributed computing with R

Karim Chine, Cloud Era Ltd

Abstract: To solve heavily computational problems, there is a need to use many engines in parallel. Several tools are available but they are difficult to install and beyond the technical skills of most scientists. Elastic-R solves this problem. From within a main R session and without installing any extra toolkits/packages, it becomes possible to create logical links to remote R/Scilab engines either by creating new processes or by connecting to existing ones on Grids/clouds. Logical links are variables that allow the R/Scilab user to interact with the remote engines. rlink.console, rlink.get, rlink.put allow the user to respectively submit R commands to the R/Scilab worker referenced by the rlink, retrieve a variable from the R/Scilab worker's workspace into the main R workspace and push a variable from the main R workspace to the worker's workspace. All the functions can be called in synchronous or asynchronous mode. Several rlinks referencing R/Scilab engines running at any locations can be used to create a logical cluster which enables to use several R/Scilab engines in a coordinated way. For example, the cluster.apply function uses the workers belonging to a logical cluster in parallel to apply a function to a large scale R data. When used in the cloud, the new functions enable scientists to leverage the elasticity of the Infrastructure-as-a-Service to control any number of R engines in parallel.

Karim Chine, "Open Science in the Cloud: Towards a Universal Platform for Scientific and Statistical Computing", Chapter 19 in "Handbook of Cloud Computing", Springer, 2010 (in Press)