sfCluster/snowfall: Managing parallel execution of R programs on a compute cluster

Jochen Knaus, Institute of Medical Biometry and Medical Informatics University Medical Center Freiburg, Germany

Modern bioinformatics applications require a huge amount of computing resources. To adress these, techniques such as MPI, available through the R packages `Rmpi` and `snow`, allow the bundling of single machines into compute clusters. However, management of cluster resources has to be performed manually, resulting in problems, when several, potentially unexperienced users access the same cluster pool.

As a solution for this problem we developed *sfCluster* and the `snowfall` R package based on the `snow` package and LAM/MPI. Both are designed for easy and safe usage, hiding cluster setup and internals from end users, who only see a clean `snow`-like API.

sfCluster is a Unix tool for management of parallel R programs, which assigns resources dynamically in a reasonable way, sets up the LAM cluster and monitors the execution of the parallel R program as well as controlling the cluster itself.

sfCluster features various execution modes: it can run an R interactive shell, raw batch mode or a visual monitoring mode, which allows process and logfile control during runtime directly on the terminal using Curses. Memory observation, process control and cluster session shutdowns even work if the LAM cluster itself died or some machines went offline or network problems occurred.

`snowfall` is the corresponding R package, which connects to *sfCluster*, but can also be used without it. In contrast to the `snow` package it provides easy switching between sequential and parallel execution, which eases development on machines without cluster environment. The package also features basic intermediate saving of results (with restore), so not all results are lost in case of a cluster stop.

The use of these advanced tools will be illustrated with application scenarios from our department, where several users can now perform demanding bioinformatics simulation studies at the same time.