CRAN Task View: High-Performance and Parallel Computing with R

Maintainer: Dirk Eddelbuettel
Contact: Dirk.Eddelbuettel at R-project.org
Version: 2013-12-18

This CRAN task view contains a list of packages, grouped by topic, that are useful for high-performance computing (HPC) with R. In this context, we are defining 'high-performance computing' rather loosely as just about anything related to pushing R a little further: using compiled code, parallel computing (in both explicit and implicit modes), working with large objects as well as profiling.

Unless otherwise mentioned, all packages presented with hyperlinks are available from CRAN, the Comprehensive R Archive Network.

Several of the areas discussed in this Task View are undergoing rapid change. Please send suggestions for additions and extensions for this task view to the task view maintainer .

Suggestions and corrections by Achim Zeileis, Markus Schmidberger, Martin Morgan, Max Kuhn, Tomas Radivoyevitch, Jochen Knaus, Tobias Verbeke, Hao Yu, David Rosenberg, Marco Enea, Ivo Welch, Jay Emerson, Wei-Chen Chen, Bill Cleveland, Ross Boylan, and Ramon Diaz-Uriarte (as well as others I may have forgotten to add here) are gratefully acknowledged.

Direct support in R started with release 2.14.0 which includes a new package **parallel** incorporating (slightly revised) copies of packages multicore and snow. Some types of clusters are not handled directly by the base package 'parallel'. However, and as explained in the package vignette, the parts of parallel which provide snow -like functions will accept snow clusters including MPI clusters.
The **parallel** package also contains support for multiple RNG streams following L'Ecuyer et al (2002), with support for both mclapply and snow clusters.
The version released for R 2.14.0 contains base functionality: higher-level convenience functions are planned for later R releases.

Parallel computing: Explicit parallelism

- Several packages provide the communications layer required for parallel computing. The first package in this area was rpvm by Li and Rossini which uses the PVM (Parallel Virtual Machine) standard and libraries. rpvm is no longer actively maintained, but available from its CRAN archive directory.
- In recent years, the alternative MPI (Message Passing Interface) standard has become the de facto standard in parallel computing. It is supported in R via the Rmpi by Yu. Rmpi package is mature yet actively maintained and offers access to numerous functions from the MPI API, as well as a number of R-specific extensions. Rmpi can be used with the LAM/MPI, MPICH / MPICH2, Open MPI, and Deino MPI implementations. It should be noted that LAM/MPI is now in maintenance mode, and new development is focussed on Open MPI.
- The pbdMPI package provides S4 classes to directly interface MPI in order to support the Single Program/Multiple Data (SPMD) parallel programming style which is particularly useful for batch parallel execution. The pbdSLAP builds on this and uses scalable linear algebra packages (namely BLACS, PBLAS, and ScaLAPACK) in double precision based on ScaLAPACK version 2.0.2. The pbdBASE builds on these and provides the core classes and methods for distributed data types upon which the pbdDMAT builds to provide distributed dense matrices for "Programming with Big Data". The pbdNCDF4 package permits multiple processes to write to the same file (without

manual synchronization) and supports terabyte-sized files. The pbdDEMO package provides examples for these packages, and a detailed vignette. The pbdPROF package profiles MPI communication SPMD code via MPI profiling libraries, such as fpmpi, mpiP, or TAU.
- An alternative is provided by the nws (NetWorkSpaces) packages from REvolution Computing. It is the successor to the earlier LindaSpaces approach to parallel computing, and is implemented on top of the Twisted networking toolkit for Python.
- The snow (Simple Network of Workstations) package by Tierney et al. can use PVM, MPI, NWS as well as direct networking sockets. It provides an abstraction layer by hiding the communications details. The snowFT package provides fault-tolerance extensions to snow.
- The snowfall package by Knaus provides a more recent alternative to snow. Functions can be used in sequential or parallel mode.
- The biopara package by Lazar and Schoenfeld offers socket-based parallel execution with some support for load-balancing and fault-tolerance.
- The foreach package allows general iteration over elements in a collection without the use of an explicit loop counter. Using foreach without side effects also facilitates executing the loop in parallel which is possible via the doMC (using multicore on single workstations), doSNOW (using snow, see above), doMPI (using Rmpi) packages and doRedis (using rredis) packages.
- The bigrf package provides a Random Forests implementation with support for parellel execution and large memory.

Parallel computing: Implicit parallelism

- The pnmath package by Tierney (link) uses the Open MP parallel processing directives of recent compilers (such gcc 4.2 or later) for implicit parallelism by replacing a number of internal R functions with replacements that can make use of multiple cores --- without any explicit requests from the user. The alternate pnmath0 package offers the same functionality using Pthreads for environments in which the newer compilers are not available. Similar functionality is expected to become integrated into R 'eventually'.
- The romp package by Jamitzky was presented at useR! 2008 (slides) and offers another interface to Open MP using Fortran. The code is still pre-alpha and available from the Google Code project romp. An R-Forge project romp was initiated but there is no package, yet.
- The fork package by Warnes provides R-equivalents to low-level Unix system functions like fork, signal, wait, kill and exit in order to spawn sub-processes for parallel execution.
- The multicore package by Urbanek provides a way of running parallel computations in R on machines with multiple cores or CPUs by making use of operating system facilities.
- The R/parallel package by Vera, Jansen and Suppi offers a C++-based master-slave dispatch mechanism for parallel execution (link)
- The Rdsm package provides a threads-like parallel computing environment, both on multicore machine and across the network by providing facilities inspired from distributed shared memory programming.
- The mchof package provides convenient, consistent parallel implementations of several commonly used higher-order functions found in base R such as Filter or Reduce.
- The RhpcBLASctl detects the number of available BLAS cores, and permits explicit selection of the number of cores.
- The Rhpc permits `*apply()` style dispatch via MPI.

Parallel computing: Grid computing

- The multiR package by Grose was presented at useR! 2008 but has not been released. It may offer a snow-style framework on a grid computing platform.
- The biocep-distrib project by Chine offers a Java-based framework for local, Grid, or Cloud

computing. It is under active development.
- The xgrid package provides functions to use Apple Xgrid clusters from within R.

Parallel computing: Hadoop

- The RHIPE package, started by Saptarshi Guha and now developed by a core team via GitHub, provides an interface between R and Hadoop for analysis of large complex data wholly from within R using the Divide and Recombine approach to big data. (link)
- The rmr package by Revolution Analytics also provides an interface between R and Hadoop for a Map/Reduce programming framework. (link)
- A related package, segue package by Long, permits easy execution of embarassingly parallel task on Elastic Map Reduce (EMR) at Amazon. (link)
- The RProtoBuf package provides an interface to Google's language-neutral, platform-neutral, extensible mechanism for serializing structured data. This package can be used in R code to read data streams from other systems in a distributed MapReduce setting where data is serialized and passed back and forth between tasks.
- The HistogramTools package provides a number of routines useful for the construction, aggregation, manipulation, and plotting of large numbers of Histograms such as those created by Mappers in a MapReduce application.

Parallel computing: Random numbers

- Random-number generators for parallel computing are available via the rsprng package by Li, and the rlecuyer package by Sevcikova and Rossini.
- The doRNG package provides functions to perform reproducible parallel foreach loops, using independent random streams as generated by the package rstream, suitable for the different foreach backends.

Parallel computing: Resource managers and batch schedulers

- Job-scheduling toolkits permit management of parallel computing resources and tasks. The slurm (Simple Linux Utility for Resource Management) set of programs (written by a consortium led by Lawrence Livermore Labs) works well with MPI. (link)
- The Condor toolkit (link) from the University of Wisconsin-Madison has been used with R as described in this R News article .
- The sfCluster package by Knaus can be used with snowfall. (link) but is currently limited to LAM/MPI.
- The batch package by Hoffmann can launch parallel computing requests onto a cluster and gather results.
- The BatchJobs package provides Map, Reduce and Filter variants to manage R jobs and their results on batch computing systems like PBS/Torque, LSF and Sun Grid Engine. Multicore and SSH systems are also supported. The BatchExperiments package extends it with an abstraction layer for running statistical experiments.

Parallel computing: Applications

- The caret package by Kuhn can use various frameworks (MPI, NWS etc) to parallelized cross-validation and bootstrap characterizations of predictive models.
- The maanova package on Bioconductor by Wu can use snow and Rmpi for the analysis of micro-array experiments.
- The pvclust package by Suzuki and Shimodaira can use snow and Rmpi for hierarchical clustering

via multiscale bootstraps; and the scaleboot package by Shimodaira can use pvclust, snow and Rmpi for computing approximately unbiased p-values via multiscale bootstraps.
- The tm package by Feinerer can use snow and Rmpi for parallelized text mining.
- The varSelRF package by Diaz-Uriarte can use snow and Rmpi for parallelized use of variable selection via random forests.
- The bcp package by Erdman and Emerson for the Bayesian analysis of change points can use foreach for parallelized operations.
- The multtest package by Pollard et al. on Bioconductor can use snow, Rmpi or rpvm for resampling-based testing of multiple hypothesis.
- The GAMBoost package by Binder for `glm` and `gam` model fitting via boosting using b-splines, the Geneland package by Estoup, Guillot and Santos for structure detection from multilocus genetic data, the Matching package by Sekhon for multivariate and propensity score matching, the STAR package by Pouzat for spike train analysis, the bnlearn package by Scutari for bayesian network structure learning, the latentnet package by Krivitsky and Handcock for latent position and cluster models, the lga package by Harrington for linear grouping analysis, the peperr package by Porzelius and Binder for parallised estimation of prediction error, the orloca package by Fernandez-Palacin and Munoz-Marquez for operations research locational analysis, the rgenoud package by Mebane and Sekhon for genetic optimization using derivatives the affyPara package by Schmidberger, Vicedo and Mansmann for parallel normalization of Affymetrix microarrays, and the puma package by Pearson et al. which propagates uncertainty into standard microarray analyses such as differential expression all can use snow for parallelized operations using either one of the MPI, PVM, NWS or socket protocols supported by snow.
- The bugsparallel package uses Rmpi for distributed computing of multiple MCMC chains using WinBUGS.
- The partDSA package uses nws for generating a piecewise constant estimation list of increasingly complex predictors based on an intensive and comprehensive search over the entire covariate space.
- The dclone package provides a global optimization approach and a variant of simulated annealing which exploits Bayesian MCMC tools to get MLE point estimates and standard errors using low level functions for implementing maximum likelihood estimating procedures for complex models using data cloning and Bayesian Markov chain Monte Carlo methods with support for JAGS, WinBUGS and OpenBUGS; parallel computing is supported via the snow package.
- The pmclust package utilizes unsupervised model-based clustering for high dimensional (ultra) large data. The package uses pbdMPI to perform a parallel version of the EM algorithm for finite mixture Gaussian models.
- The harvestr package provides helper functions for (reproducible) simulations.
- Nowadays, many packages can use the facilities offered by the **parallel** (or multicore) package. One example is pls

Parallel computing: GPUs

- The gputools package by Buckner and Seligman provides several common data-mining algorithms which are implemented using a mixture of nVidia's CUDA langauge and cublas library. Given a computer with an nVidia GPU these functions may be substantially more efficient than native R routines. The rpud package provides an optimised distance metric for NVidia-based GPUs.
- The cudaBayesreg package by da Silva implements the `rhierLinearModel` from the bayesm package using nVidia's CUDA langauge and tools to provide high-performance statistical analysis of fMRI voxels.
- The rgpu package (see below for link) aims to speed up bioinformatics analysis by using the GPU.
- The magma package provides an interface to the hybrid GPU/CPU library Magma (see below for link).
- The gcbd package implements a benchmarking framework for BLAS and GPUs (using gputools).

- The OpenCL package provides an interface from R to OpenCL permitting hardware- and vendor neutral interfaces to GPU programming.
- The WideLM package use CUDA (4.1 or greater) to fit many 'skinny' regression models simultaneously from a single data set.
- The HiPLARM package provide High-Performance Linear Algebra for R using multi-core and/or GPU support using the PLASMA / MAGMA libraries from UTK, CUDA, and accelerated BLAS.
- The permGPU package computes permutation resampling inference in the context of RNA microarray studies on the GPU, it uses CUDA (>= 4.5)

Large memory and out-of-memory data

- The biglm package by Lumley uses incremental computations to offer `lm()` and `glm()` functionality to data sets stored outside of R's main memory.
- The ff package by Adler et al. offers file-based access to data sets that are too large to be loaded into memory, along with a number of higher-level functions.
- The bigmemory package by Kane and Emerson permits storing large objects such as matrices in memory (as well as via files) and uses external pointer objects to refer to them. This permits transparent access from R without bumping against R's internal memory limits. Several R processes on the same computer can also share big memory objects.
- A large number of database packages, and database-alike packages (such as sqldf by Grothendieck and data.table by Dowle) are also of potential interest but not reviewed here.
- The HadoopStreaming package provides a framework for writing map/reduce scripts for use in Hadoop Streaming; it also facilitates operating on data in a streaming fashion which does not require Hadoop.
- The speedglm package permits to fit (generalised) linear models to large data. For in-memory data sets, speedlm() or speedglm() can be used along with update.speedlm() which can update fitted models with new data. For out-of-memory data sets, shglm() is available; it works in the presence of factors and can check for singular matrices.
- The biglars package by Seligman et al can use the ff to support large-than-memory datasets for least-angle regression, lasso and stepwise regression.
- The bigrf package provides a Random Forests implementation with support for parellel execution and large memory.
- The MonetDB.R package allows R to access the MonetDB column-oriented, open source database system as a backend.

Easier interfaces for Compiled code

- The inline package by Sklyar et al eases adding code in C, C++ or Fortran to R. It takes care of the compilation, linking and loading of embeded code segments that are stored as R strings.
- The Rcpp package by Eddelbuettel and Francois offers a number of C++ clases that makes transferring R objects to C++ functions (and back) easier, and the RInside package by the same authors allows easy embedding of R itself into C++ applications for faster and more direct data transfer.
- The rJava package by Urbanek provides a low-level interface to Java similar to the `.Call()` interface for C and C++.

Profiling tools

- The profr package by Wickham can visualize output from the `Rprof` interface for profiling.
- The proftools package by Tierney can also be used to analyse profiling output.

CRAN packages:

- batch
- BatchExperiments
- BatchJobs
- bayesm
- bcp
- biglars
- biglm
- bigmemory
- bigrf
- biopara
- bnlearn
- caret
- cudaBayesreg
- data.table
- dclone
- doMC
- doMPI
- doRedis
- doRNG
- doSNOW
- ff
- foreach
- fork
- GAMBoost
- gcbd
- Geneland
- gputools
- HadoopStreaming
- harvestr
- HiPLARM
- HistogramTools
- inline
- latentnet
- lga
- magma
- Matching
- mchof
- MonetDB.R
- multicore
- nws
- OpenCL
- orloca
- partDSA
- pbdBASE
- pbdDEMO
- pbdDMAT
- pbdMPI
- pbdNCDF4
- pbdPROF

- pbdSLAP
- peperr
- permGPU
- pls
- pmclust
- profr
- proftools
- pvclust
- Rcpp
- Rdsm
- rgenoud
- Rhpc
- RhpcBLASctl
- RInside
- rJava
- rlecuyer
- Rmpi (core)
- RProtoBuf
- rpud
- rredis
- rsprng
- scaleboot
- snow (core)
- snowfall
- snowFT
- speedglm
- sqldf
- STAR
- tm
- varSelRF
- WideLM
- xgrid

Related links:

- HPC computing notes by Luke Tierney for HPC class at University of Iowa
- Mailing List: R Special Interest Group High Performance Computing
- Schmidberger, Morgan, Eddelbuettel, Yu, Tierney and Mansmann (2009) paper on 'State of the Art in Parallel Computing with R'
- Luke Tierney's code directory for pnmath and pnmath0
- R-Forge Project: biocep-distrib
- Bioconductor Package: affyPara
- Bioconductor Package: maanova
- Bioconductor Package: multtest
- Bioconductor Package: puma
- Google Code Project: romp
- Google Code Project: bugsparallel
- Slurm project at Lawrence Livermore National Laboratory
- Condor project at University of Wisconsin-Madison
- Parallel Computing in R with sfCluster/snowfall
- Wikipedia: Message Passing Interface (MPI)

- Wikipedia: Parallel Virtual Machine (PVM)
- Slides from Introduction to High-Performance Computing with R tutorial help in Nov 2009 at the Institute for Statistical Mathematics, Tokyo, Japan
- rgpu project at nbic.nl
- Magma: Matrix Algebra on GPU and Multicore architectures
- Parallel R: Data Analysis in the Distributed World"
- High Performance Statistical Computing for Data Intensive Research
- Rth: Parallel R through Thrust
- Programming with Big Data in R
- RHIPE