Introduction to *BatchtoolsParam*

Nitesh Turaga[1]***, Martin Morgan***[2]

[1]Nitesh.Turaga@RoswellPark.org

[2]Martin.Morgan@RoswellPark.org

Edited: March 22, 2018; Compiled: April 17, 2024

Contents

1 Introduction

The `BatchtoolsParam` class is an interface to the *batchtools* package from within *BiocParallel*, for computing on a high performance cluster such as SGE, TORQUE, LSF, SLURM, OpenLava.

2 Quick start

This example demonstrates the easiest way to launch a 100000 jobs using batchtools. The first step involves creating a `BatchtoolsParam` class. You can compute using 'bplapply' and then the result is stored.

```
library(BiocParallel)

## Pi approximation
piApprox <- function(n) {
    nums <- matrix(runif(2 * n), ncol = 2)
    d <- sqrt(nums[, 1]^2 + nums[, 2]^2)
    4 * mean(d <= 1)
}

piApprox(1000)
```

```
## [1] 3.088
```

```
## Apply piApprox over
param <- BatchtoolsParam()
result <- bplapply(rep(10e5, 10), piApprox, BPPARAM=param)
mean(unlist(result))
```

```
## [1] 3.141275
```

3 *BatchtoolsParam* interface

The `BatchtoolsParam` interface allows intuitive usage of your high performance cluster with *BiocParallel*.

The `BatchtoolsParam` class allows the user to specify many arguments to customize their jobs. Applicable to clusters with formal schedulers.

- `workers` The number of workers used by the job.
- `cluster` We currently support, SGE, SLURM, LSF, TORQUE and OpenLava. The 'cluster' argument is supported only if the R environment knows how to find the job scheduler. Each cluster type uses a template to pass the job to the scheduler. If the template is not given we use the default templates as given in the 'batchtools' package. The cluster can be accessed by 'bpbackend(param)'.
- `registryargs` The 'registryargs' argument takes a list of arguments to create a new job registry for you `BatchtoolsParam`. The job registry is a data.table which stores all the required information to process your jobs. The arguments we support for registryargs are:

  `file.dir` Path where all files of the registry are saved. Note that some templates do not handle relative paths well. If nothing is given, a temporary directory will be used in your current working directory.

  `work.dir` Working directory for R process for running jobs.

  `packages` Packages that will be loaded on each node.

  `namespaces` Namespaces that will be loaded on each node.

  `source` Files that are sourced before executing a job.

  `load` Files that are loaded before executing a job.

```
registryargs <- batchtoolsRegistryargs(
    file.dir = "mytempreg",
    work.dir = getwd(),
    packages = character(0L),
    namespaces = character(0L),
    source = character(0L),
    load = character(0L)
)
param <- BatchtoolsParam(registryargs = registryargs)
param
```

```
## class: BatchtoolsParam
##   bpisup: FALSE; bpnworkers: 4; bptasks: 0; bpjobname: BPJOB
##   bplog: FALSE; bpthreshold: INFO; bpstopOnError: TRUE
```

```
##   bpRNGseed: NA; bptimeout: NA; bpprogressbar: FALSE
##   bpexportglobals: TRUE; bpexportvariables: TRUE; bpforceGC: FALSE
##   bpfallback: TRUE
##   bplogdir: NA
##   bpresultdir: NA
##   cluster type: multicore
##   template: NA
##   registryargs:
##     file.dir: mytempreg
##     work.dir: /tmp/Rtmpg8CuaW/Rbuild6f49f1901f12f/BiocParallel/vignettes
##     packages: character(0)
##     namespaces: character(0)
##     source: character(0)
##     load: character(0)
##     make.default: FALSE
##   saveregistry: FALSE
##   resources:
```

- `resources` A named list of key-value pairs to be subsituted into the template file; see `?batchtools::submitJobs`.
- `template` The template argument is unique to the `BatchtoolsParam` class. It is required by the job scheduler. It defines how the jobs are submitted to the job scheduler. If the template is not given and the cluster is chosen, a default template is selected from the batchtools package.
- `log` The log option is logical, TRUE/FALSE. If it is set to TRUE, then the logs which are in the registry are copied to directory given by the user using the `logdir` argument.
- `logdir` Path to the logs. It is given only if `log=TRUE`.
- `resultdir` Path to the directory is given when the job has files to be saved in a directory.

4 Defining templates

The job submission template controls how the job is processed by the job scheduler on the cluster. Obviously, the format of the template will differ depending on the type of job scheduler. Let's look at the default SLURM template as an example:

```
fname <- batchtoolsTemplate("slurm")

## using default 'slurm' template in batchtools.

cat(readLines(fname), sep="\n")

## #!/bin/bash
##
## ## Job Resource Interface Definition
## ##
## ## ntasks [integer(1)]:       Number of required tasks,
## ##                            Set larger than 1 if you want to further parallelize
## ##                            with MPI within your job.
## ## ncpus [integer(1)]:        Number of required cpus per task,
## ##                            Set larger than 1 if you want to further parallelize
```

```
## ##                            with multicore/parallel within each task.
## ## walltime [integer(1)]:     Walltime for this job, in seconds.
## ##                            Must be at least 60 seconds for Slurm to work properly.
## ## memory   [integer(1)]:     Memory in megabytes for each cpu.
## ##                            Must be at least 100 (when I tried lower values my
## ##                            jobs did not start at all).
## ##
## ## Default resources can be set in your .batchtools.conf.R by defining the variable
## ## 'default.resources' as a named list.
##
## <%
## # relative paths are not handled well by Slurm
## log.file = fs::path_expand(log.file)
## -%>
##
##
## #SBATCH --job-name=<%= job.name %>
## #SBATCH --output=<%= log.file %>
## #SBATCH --error=<%= log.file %>
## #SBATCH --time=<%= ceiling(resources$walltime / 60) %>
## #SBATCH --ntasks=1
## #SBATCH --cpus-per-task=<%= resources$ncpus %>
## #SBATCH --mem-per-cpu=<%= resources$memory %>
## <%= if (!is.null(resources$partition)) sprintf(paste0("#SBATCH --partition='", resources$partition, "'")) %>
## <%= if (array.jobs) sprintf("#SBATCH --array=1-%i", nrow(jobs)) else "" %>
##
## ## Initialize work environment like
## ## source /etc/profile
## ## module add ...
##
## ## Export value of DEBUGME environemnt var to slave
## export DEBUGME=<%= Sys.getenv("DEBUGME") %>
##
## <%= sprintf("export OMP_NUM_THREADS=%i", resources$omp.threads) -%>
## <%= sprintf("export OPENBLAS_NUM_THREADS=%i", resources$blas.threads) -%>
## <%= sprintf("export MKL_NUM_THREADS=%i", resources$blas.threads) -%>
##
## ## Run R:
## ## we merge R output with stdout from SLURM, which gets then logged via --output option
## Rscript -e 'batchtools::doJobCollection("<%= uri %>")'
```

The `<%= =>` blocks are automatically replaced by the values of the elements in the `resources` argument in the `BatchtoolsParam` constructor. Failing to specify critical parameters properly (e.g., wall time or memory limits too low) will cause jobs to crash, usually rather cryptically. We suggest setting parameters explicitly to provide robustness to changes to system defaults. Note that the `<%= =>` blocks themselves do not usually need to be modified in the template.

The part of the template that is most likely to require explicit customization is the last line containing the call to `Rscript`. A more customized call may be necessary if the R installation is not standard, e.g., if multiple versions of R have been installed on a cluster. For example, one might use instead:

```
echo 'batchtools::doJobCollection("<%= uri %>")' |\
    ArbitraryRcommand --no-save --no-echo
```

If such customization is necessary, we suggest making a local copy of the template, modifying it as required, and then constructing a `BiocParallelParam` object with the modified template using the `template` argument. However, we find that the default templates accessible with `batchtoolsTemplate` are satisfactory in most cases.

5 Use cases

As an example for a BatchtoolParam job being run on an SGE cluster, we use the same `piApprox` function as defined earlier. The example runs the function on 5 workers and submits 100 jobs to the SGE cluster.

Example of SGE with minimal code:

```
library(BiocParallel)

## Pi approximation
piApprox <- function(n) {
    nums <- matrix(runif(2 * n), ncol = 2)
    d <- sqrt(nums[, 1]^2 + nums[, 2]^2)
    4 * mean(d <= 1)
}

template <- system.file(
    package = "BiocParallel",
    "unitTests", "test_script", "test-sge-template.tmpl"
)
param <- BatchtoolsParam(workers=5, cluster="sge", template=template)

## Run parallel job
result <- bplapply(rep(10e5, 100), piApprox, BPPARAM=param)
```

Example of SGE demonstrating some of `BatchtoolsParam` methods.

```
library(BiocParallel)

## Pi approximation
piApprox <- function(n) {
    nums <- matrix(runif(2 * n), ncol = 2)
    d <- sqrt(nums[, 1]^2 + nums[, 2]^2)
    4 * mean(d <= 1)
}

template <- system.file(
    package = "BiocParallel",
    "unitTests", "test_script", "test-sge-template.tmpl"
)
param <- BatchtoolsParam(workers=5, cluster="sge", template=template)

## start param
```

```
bpstart(param)

## Display param
param

## To show the registered backend
bpbackend(param)

## Register the param
register(param)

## Check the registered param
registered()

## Run parallel job
result <- bplapply(rep(10e5, 100), piApprox)

bpstop(param)
```

6 sessionInfo()

```
toLatex(sessionInfo())
```

- R version 4.4.0 beta (2024-04-15 r86425), `x86_64-pc-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_GB`, `LC_COLLATE=C`, `LC_MONETARY=en_US.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=en_US.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Time zone: `America/New_York`
- TZcode source: `system (glibc)`
- Running under: `Ubuntu 22.04.4 LTS`
- Matrix products: default
- BLAS: `/home/biocbuild/bbs-3.19-bioc/R/lib/libRblas.so`
- LAPACK: `/usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.10.0`
- Base packages: base, datasets, grDevices, graphics, methods, stats, stats4, utils
- Other packages: Biobase 2.63.1, BiocGenerics 0.49.1, BiocParallel 1.37.1, BiocStyle 2.31.0, Biostrings 2.71.5, GenomeInfoDb 1.39.14, GenomicAlignments 1.39.5, GenomicRanges 1.55.4, IRanges 2.37.1, MatrixGenerics 1.15.0, RNAseqData.HNRNPC.bam.chr14 0.41.0, Rsamtools 2.19.4, S4Vectors 0.41.6, SummarizedExperiment 1.33.3, XVector 0.43.1, matrixStats 1.3.0
- Loaded via a namespace (and not attached): BiocManager 1.30.22, DelayedArray 0.29.9, GenomeInfoDbData 1.2.12, Matrix 1.7-0, R6 2.5.1, S4Arrays 1.3.7, SparseArray 1.3.5, UCSC.utils 0.99.7, abind 1.4-5, backports 1.4.1, base64url 1.4, batchtools 0.9.17, bitops 1.0-7, bookdown 0.39, brew 1.0-10, bslib 0.7.0, cachem 1.0.8, checkmate 2.3.1, cli 3.6.2, codetools 0.2-20, compiler 4.4.0,

crayon 1.5.2, data.table 1.15.4, debugme 1.1.0, digest 0.6.35, evaluate 0.23,
fansi 1.0.6, fastmap 1.1.1, fs 1.6.3, glue 1.7.0, grid 4.4.0, highr 0.10, hms 1.1.3,
htmltools 0.5.8.1, httr 1.4.7, jquerylib 0.1.4, jsonlite 1.8.8, knitr 1.46, lattice 0.22-6,
lifecycle 1.0.4, magrittr 2.0.3, parallel 4.4.0, pillar 1.9.0, pkgconfig 2.0.3,
prettyunits 1.2.0, progress 1.2.3, rappdirs 0.3.3, rlang 1.1.3, rmarkdown 2.26,
sass 0.4.9, snow 0.4-4, stringi 1.8.3, tibble 3.2.1, tools 4.4.0, utf8 1.2.4, vctrs 0.6.5,
withr 3.0.0, xfun 0.43, yaml 2.3.8, zlibbioc 1.49.3