Guide to the **pbdPROF** Package

Wei-Chen Chen[1], Drew Schmidt[2], Gaurav Sehrawat[3], Pragneshkumar Patel[2], and George Ostrouchov[1,2]

[1]Computer Science and Mathematics Division,
Oak Ridge National Laboratory,
Oak Ridge, TN, USA

[2]Remote Data Analysis and Visualization Center
University of Tennessee,
Knoxville, TN, USA

[3]Jaypee Institute of Information Technology
Uttar Pradesh, India

September 21, 2013

Contents

This publication was typeset using LaTeX.

Acknowledgement

Chen and Ostrouchov were supported in part by the project "Visual Data Exploration and Analysis of Ultra-large Climate Data" funded by U.S. DOE Office of Science under Contract No. DE-AC05-00OR22725. Ostrouchov, Schmidt, and Patel were supported in part by the project "NICS Remote Data Analysis and Visualization Center" funded by the Office of Cyberinfrastructure of the U.S. National Science Foundation under Award No. ARRA-NSF-OCI-0906324 for NICS-RDAV center.

Sehrawat was generously supported by Google for Google Summer of Code 2013.

Warning: The findings and conclusions in this article have not been formally disseminated by the U.S. Department of Energy and should not be construed to represent any determination or policy of University, Agency, and National Laboratory.

This document is written to explain the main functions of **pbdPROF** (Chen *et al.*, 2013), version 0.1-0. Every effort will be made to ensure future versions are consistent with these instructions, but features in later versions may not be explained in this document.

Information about the functionality of this package, and any changes in future versions can be found on website: "Programming with Big Data in R" at `http://r-pbd.org/`.

1 Introduction

The goal of **pbdPROF** is to utilize external MPI profiling libraries, such as **fpmpi** (Gropp, 2000), **mpiP** (Vetter and McCracken, 2001), or **TAU** (Shende and Malony, 2006), to profile parallel R code and understand hidden MPI communications between processors. The number of communications, sizes of messages, times, and types of functions calls all affect program performance, and so having these measurements can greatly aid in debugging and algorithm design. These MPI profiling libraries are able to hijack calls to MPI functions and then capture the profiling information (as described above), all without disturbing the execution of the original program.

The current main features of **pbdPROF** include:

1. providing linking information to pbdR (Ostrouchov *et al.*, 2012) and other MPI-using R packages
2. output profiling information associated with MPI calls,
3. parsing and summarizing profiling information, and
4. support several MPI profiling libraries.

2 Installation

2.1 System Requirements

The **pbdPROF** package requires an MPI installation, such as OpenMPI or MPICH2. Additionally, the package is basically useless without some kind of MPI-using R package, such as **pbdMPI** (Chen *et al.*, 2012a) or **Rmpi** (Yu, 2002). For information regarding how to install MPI or **pbdMPI**, please see the **pbdMPI** vignette (Chen *et al.*, 2012b) or the pbdR website `http://r-pbd.org/install`.

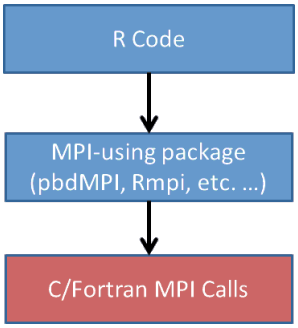
2.2 The Big Picture

Before pressing on, let us stop to take a moment and understand the "big picture" here. The following sections will contain *more than sufficient* detail, to the point where it would be easy to lose sight of the proverbial forest for the trees.

For the remainder of this document, we will be providing information for two fairly distinct groups of people: R-level MPI package developers, and C/Fortran-level MPI package developers. If you are in the former category, then the use of this package is a bit simpler for you. All you need to do is get **pbdPROF**

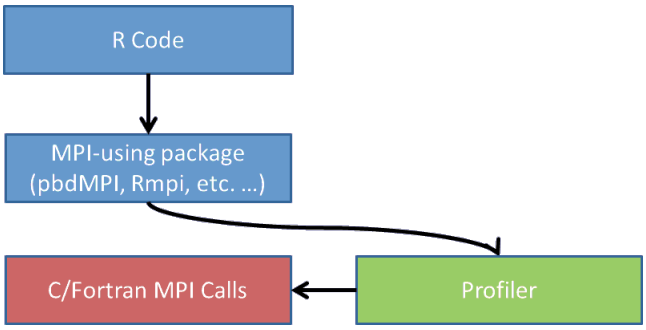
installed and reinstall your MPI-using package of choice (**pbdMPI**, **Rmpi**, etc. ...). Each package that directly uses MPI (packages produced by developers in the latter category) will have to explicitly support **pbdPROF** (or the reader will have to get his/her hands dirty in another developer's makefiles — an unpleasant business). It is worth nothing here that there are instructions in this document for how a developer of the second kind could explicitly add **pbdPROF** support to his/her package.

So why the need to reinstall things? It boils down to how the profilers actually work. Under normal circumstances, a user writes some R code from an MPI-using package (e.g., `allreduce(x)` from **pbdMPI**, `mpi.allreduce(x, type=2)` from **Rmpi**, etc. ...). This then makes a call to some C or Fortran code



Without a Profiler

which directly interfaces with MPI. You can see this pictures in Figure 1. When you use a profiler, you instead hijack the calls to MPI from the C/Fortran code so that some metadata can be stored about MPI usage. This process is represented in Figure 2. Hopefully it should be clear what, and when, something



With the Profiler

should be reinstalled. For the sake of completion, we summarize the possibilities below:

To *enable* MPI profiling:

1. install **pbdPROF**
2. reinstall an MPI-using package and link it with **pbdPROF**
3. write and execute your MPI-using R code as normal
4. use the **pbdPROF** utilities `read.prof()`, `plot()`, etc. for interpreting profiling results

To *disable* MPI profiling:

1. reinstall any MPI-using package that was linked it with **pbdPROF**, and this time *do not* link with **pbdPROF**

2.3 Choice of Profiler

The **pbdPROF** package currently uses the **fpmpi** library by default. More explicitly, a source copy of **fpmpi** is located at `pbdPROF/src/fpmpi` of the **pbdPROF** source. If this profiler is used, static library will be built and placed in `pbdPROF/lib/libfpmpi.a` of the **pbdPROF** install directory. However, external profiling libraries such as **mpiP**, **TAU**, or even **fpmpi** can be also linked with **pbdPROF** by passing a suitable `--configure-args` argument during an installation via `R CMD INSTALL`. We will explain this procedure in depth in Section 2.4 using an external **fpmpi** as an example, but we will leave some keys steps for **mpiP** and **TAU** for future Sections.

While it is possible to link with other profiling libraries, at the time of writing (for version 0.1-0), we only support **fpmpi**. We anticipate full support of **mpiP** and **TAU** for the next version of this package.

Regardless of whether **fpmpi**, **mpiP**, or **TAU** is used, we strongly recommend adding `CPPFLAGS="-fPIC"` at the `configure` step.

2.4 fpmpi

We can install **pbdPROF** using the internal **fpmpi** library via

Shell Command

```
R CMD INSTALL pbdPROF_0.1-0.tar.gz
```

By default, this compiles `pbdPROF/src/fpmpi/*` of the **pbdPROF** source, generates a static library `libfpmpi.a`, and installs the library to `pbdPROF/lib/` of the **pbdPROF** install. No shared library is generated or needed, so the directory `pbdPROF/libs/` is empty, i.e., there is no need to build `pbdPROF.so`. The linking argument is saved in `Makeconf` and installed to `pbdPROF/etc/` for later use by other packages, such as **pbdMPI** or **Rmpi**.

However, if we choose, we can link with an external **fpmpi** library, via

Shell Command

```
R CMD INSTALL pbdPROF_0.1-0.tar.gz \
  --configure-args="--with-fpmpi='/path_to_fpmpi/lib/libfpmpi.a'"
```

or

Shell Command

```
R CMD INSTALL pbdPROF_0.1-0.tar.gz \
  --configure-args="--with-fpmpi='-L/path_to_fpmpi/lib -lfpmpi'"
```

Since **fpmpi** only builds a static library `libfpmpi.a`, there is no difference between these two installations of **pbdPROF**. This only provides the linking arguments, either `/path_to_fpmpi/lib/libfpmpi.a` or `-L/path_to_fpmpi/lib -lfpmpi`, which is saved in `Makeconf` and installed to `pbdPROF/etc/` for later use by other packages, such as **pbdMPI** or **Rmpi**.

2.4.1 Reinstall pbdMPI

Reinstall **pbdMPI** via

Shell Command

```
R CMD INSTALL pbdMPI_1.0-0.tar.gz --configure-args="--enable-pbdPROF"
```

Package developers who are directly interfacing with MPI (via C or Fortran) should note that `pbdMPI/R/get_conf.r` and `pbdMPI/R/get_lib.r` are used in `pbdMPI/configure.ac` or `pbdMPI/configure` to determine an appropriate linking flag `PROF_LDFLAGS` based on preset flags in `pbdPROF/etc/Makeconf`.

If the internal library is used in **pbdPROF**, then the path to `pbdPROF/lib/libfpmpi.a` is set in the flag `PKG_LIBS` of `pbdMPI/src/Makevars.in`. If the external library is used in **pbdPROF**, then the linking arguments either `/path_to_fpmpi/lib/libfpmpi.a` or `-L/path_to_fpmpi/lib -lfpmpi` is set in the flag `PKG_LIBS` of `pbdMPI/src/Makevars.in`. Therefore, the **pbdMPI** can be intercepted by the **fpmpi** library when MPI function calls are evoked.

No mater which library is used, internal or external, the `PROF_LDFLAGS` in `pbdMPI/etc/Makefile` provides the linking information to the profiling library. It is also used in `PKG_LIBS`, which will be exported to other pbdR packages at installation via the flag `SPMD_LDFLAGS`. Therefore there is no need for additional flags in `R CMD INSTALL` when reinstalling packages for profiling.

2.4.2 Reinstall pbdBASE

For further profiling, such as **pbdBASE** (Schmidt *et al.*, 2012), one may reinstall the package, via

Shell Command

```
R CMD INSTALL pbdBASE_0.2-2.tar.gz
```

There is no need to provide any flag since **pbdMPI/etc/Makefile** has the information and installation of **pbdBASE** already considers it. Note that since both packages (**pbdMPI** and **pbdBASE**) have MPI-using C/Fortran functions involved, it is necessary to link with **pbdPROF** in order to profile communications evoked by the package.

2.4.3 Reinstall Rmpi

Reinstall **Rmpi** via

Shell Command

```
wget https://github.com/snoweye/Rmpi_PROF/archive/master.zip
unzip master.zip
mv Rmpi_PROF-master Rmpi
find ./Rmpi -type f -perm 777 -print -exec chmod 644 {} \;
find ./Rmpi -type d -perm 777 -print -exec chmod 755 {} \;
chmod 755 ./Rmpi/configure
chmod 755 ./Rmpi/cleanup
chmod 755 ./Rmpi/inst/*.sh
R CMD build --no-resave-data Rmpi
R CMD INSTALL Rmpi_0.6-4.tar.gz --configure-args="--enable-pbdPROF'"
```

Note that 0.6-4 is not an official release of **Rmpi**. It is a modified version of 0.6-3 and it is currently available at https://github.com/snoweye/Rmpi_PROF. The authors of **Rmpi** have plans to eventually incorporate these changes, but this can be used as a temporary measure.

3 Test Scripts

We provide two short R scripts, one for **pbdMPI** and one for **Rmpi**, to test the installation and profiling capabilities of **pbdPROF**. If the installation is correct, then executing these examples codes should produce profiler output.

3.1 Test with pbdMPI

Below we provide sample scripts to test that the installation of **pbdPROF** was successful. For **pbdMPI**, use:

Test script for pbdMPI

```
### Save this in a file: prof_pbdMPI.r
library(pbdMPI, quiet = TRUE)
init()

set.seed(comm.rank())
x <- allreduce(rnorm(100), op = "sum")

finalize()
```

and run this code by

R Script

```
mpiexec -np 2 Rscript prof_pbdMPI.r
```

A successful output of **fpmpi** in the profiling file `fpmpi_profile.txt` may contain

```
Details for each MPI routine
                  Average of sums over all processes
                                                  % by message length
                                 (max over        0.........1........1........
                                  processes [rank])           K        M
MPI_Allreduce:
        Calls      :           2            2 [   0] 05000000050000000000000000000
        Time       :   3.61e-05     3.72e-05 [   0] 07000000030000000000000000000
        Data Sent  :         804          804 [   0]
        SyncTime   :     0.00149      0.00287 [   0] 0*0000000.000000000000000000
        By bin     : 1-4 [1,1]    [   2.5e-05,  2.72e-05] [   4.1e-05,    0.00286]
                   : 513-1024     [1,1]    [     1e-05,     1e-05] [   1.1e-05,
                      7.61e-05]
```

In this R script, one MPI C function `MPI_Allreduce` is called twice and 804 bytes are sent that a hundred of double precision (8 bytes) for 100 normal random variables, and one integer (4 bytes) for checking data type to call the corresponding S4 method.

3.2 Test with Rmpi

For **Rmpi**, use:

Test script for pbdMPI

```
### Save this in a file: prof_Rmpi.r
library(Rmpi, quiet = TRUE)
mpi.comm.dup(0, 1)

set.seed(mpi.comm.rank())
x <- mpi.allreduce(rnorm(100), type = 2, op = "sum")

mpi.quit()
```

and run this code by

R Script

```
mpiexec -np 2 Rscript prof_Rmpi.r
```

A successful output of **fpmpi** in the profiling file `fpmpi_profile.txt` could be

```
Details for each MPI routine
                  Average of sums over all processes
                                                  % by message length
                                  (max over       0.........1........1........
                                   processes [rank])           K        M
MPI_Allreduce:
        Calls     :          1            1 [   0] 000000000*000000000000000000
        Time      :   4.01e-05     4.41e-05 [   1] 000000000*000000000000000000
        Data Sent :        800          800 [   0]
        SyncTime  :    0.00103      0.00204 [   1] 000000000*000000000000000000
        By bin    : 513-1024    [1,1]   [   3.6e-05,  4.41e-05] [  2.79e-05,   0
.00204]
MPI_Comm_dup:
        Calls     :          1
        Time      :   5.81e-05
        SyncTime  :   0.000211
```

Two MPI C functions `MPI_Allreduce` and `MPI_Comm_dup` are called one time for each.

4 Profiling with fpmpi

4.1 Demo of pbdMPI

The `allreduce.r` is originally in **pbdMPI/demo/** and can be profiled by

R Script

```
mpiexec -np 2 Rscript -e "demo(allreduce,'pbdMPI',ask=F,echo=F)"
```

which will provide an output file `fpmpi_profile.txt`. Part of output is listed in the next as

```
Processes:      2
Execute time:   1.176
Timing Stats: [seconds] [min/max]       [min rank/max rank]
  wall-clock: 1.176 sec 1.171488 / 1.180277     0 / 1
        user: 0.378 sec 0.360000 / 0.396000     0 / 1
```

```
         sys: 0.07 sec  0.040000 / 0.100000     1 / 0

                  Average of sums over all processes
Routine                 Calls       Time Msg Length    %Time by message length
                                                    0.........1........1........
                                                                K        M
MPI_Allreduce       :      10   0.000118        188 06100300000000000000000000
MPI_Barrier         :      21     0.0054

Details for each MPI routine
                  Average of sums over all processes
                                                  % by message length
                                (max over         0.........1........1........
                                 processes [rank])          K        M
MPI_Allreduce:
        Calls     :         10           10 [   0] 05100400000000000000000000
        Time      :   0.000118     0.000119 [   0] 06100300000000000000000000
        Data Sent :        188          188 [   0]
        SyncTime  :   0.000312     0.000453 [   0] 07.00200000000000000000000
        By bin    : 1-4 [5,5]    [  7.01e-05,  7.01e-05] [  0.000117,  0.000343]
                  : 5-8 [1,1]    [  7.87e-06,  9.06e-06] [  9.06e-06,  9.06e-06]
                  : 33-64        [4,4]    [  3.91e-05,  4.03e-05] [  4.51e-05,
                     0.0001]
MPI_Barrier:
        Calls     :         21
        Time      :     0.0054
```

Two MPI C functions `MPI_Allreduce` and `MPI_Barrier` are evoked inside this R code. The `MPI_Allreduce` is called 10 times, span 0.000156 seconds, and 188 bytes are sent. The `MPI_Barrier` is called 21 times and span 0.00608 seconds.

4.2 Demo of pbdDMAT

The `svd.r` is originally in **pbdDMA/demo/** (Schmidt *et al.*, 2012) and can be profiled by

R Script

```
mpiexec -np 2 Rscript -e "demo(svd,'pbdDMAT',ask=F,echo=F)"
```

which will provide an output file `fpmpi_profile.txt`. Part of output is listed in the next as

```
Processes:      2
Execute time:   1.774
Timing Stats: [seconds] [min/max]       [min rank/max rank]
  wall-clock: 1.774 sec 1.766181 / 1.781962     1 / 0
        user: 0.962 sec 0.956000 / 0.968000     1 / 0
         sys: 0.046 sec 0.044000 / 0.048000     0 / 1

                  Average of sums over all processes
Routine                 Calls       Time Msg Length    %Time by message length
                                                    0.........1........1........
                                                                K        M
MPI_Allreduce       :      12   0.000108         72 06400000000000000000000000
MPI_Barrier         :       8   0.000784

Details for each MPI routine
```

```
                  Average of sums over all processes
                                                  % by message length
                                  (max over       0.........1........1........
                                   processes [rank])           K        M
MPI_Allreduce:
        Calls     :          12           12 [   0] 05500000000000000000000000
        Time      :    0.000108     0.000113 [   0] 06400000000000000000000000
        Data Sent :          72           72 [   0]
        SyncTime  :    0.000143      0.00016 [   1] 06400000000000000000000000
        By bin    : 1-4 [6,6]   [  5.44e-05,  6.91e-05] [  6.91e-05,  8.89e-05]
                  : 5-8 [6,6]   [  4.36e-05,  4.79e-05] [  5.72e-05,  7.08e-05]
MPI_Barrier:
        Calls     :           8
        Time      :    0.000784
```

Two MPI C functions `MPI_Allreduce` and `MPI_Barrier` are evoked inside this R code. The `MPI_Allreduce` is called 12 times, span 0.000108 seconds, and 72 bytes are sent. The `MPI_Barrier` is called 8 times and span 0.000784 seconds.

4.3 Demo of Rmpi

The `masterSlavePI.r` is originally in **Rmpi/demo/** and can be profiled by

R Script

```
mpiexec -np 4 Rscript -e "demo(masterslavePI,'Rmpi',ask=F,echo=F)"
```

which will provide an output file `fpmpi_profile.txt`. Part of output is listed in the next as

```
Processes:      1
Execute time:   0.05362
Timing Stats: [seconds] [min/max]       [min rank/max rank]
  wall-clock: 0.05362 sec       0.053622 / 0.053622     0 / 0
        user: 0.236 sec 0.236000 / 0.236000     0 / 0
         sys: 0.052 sec 0.052000 / 0.052000     0 / 0

                  Average of sums over all processes
Routine                 Calls       Time Msg Length    %Time by message length
                                                      0.........1........1........
                                                                K        M
MPI_Reduce          :       1   6.51e-05          8 00*00000000000000000000000000

Details for each MPI routine
                  Average of sums over all processes
                                                  % by message length
                                  (max over       0.........1........1........
                                   processes [rank])           K        M
MPI_Reduce:
        Calls     :           1            1 [   0] 00*00000000000000000000000000
        Time      :    6.51e-05     6.51e-05 [   0] 00*00000000000000000000000000
        Data Sent :           8            8 [   0]
        By bin    : 5-8 [1,1]   [  6.51e-05,  6.51e-05]
```

One MPI C function `MPI_Reduce` is evoked inside this R code. The `MPI_Reduce` is called only 1 time, span $6.51e-05$ seconds, and 8 bytes are sent. Note that there is only one processor (master in `comm=0`)

profiled by **fpmpi**, and the other three processors (slaves in `comm=1`) are not.

5 Problems with pbdPROF

5.1 Installation

Problem 1: If you have downloaded the package from github and tried to using R CMD INSTALL **pbdPROF** and you see an error similar to this

```
ERROR: 'configure' exists but is not executable -- see the 'R Installation and
    Administration Manual'
```

Solution: You have to make the configure executable which means giving it permission , which can done by

R Script

```
chmod +x configure
```

after changing the folder to package's main directory.

Problem 2: If you are using **fpmpi** (Gropp, 2000) externally and during it's installation you get an error similar to this

```
error :checking for library containing MPI_Init... (cached) no configure:
error: Could not find MPI library
```

Solution: You probably need to specify the path to MPI library using this in command line in the fpmpi main directory

R Script

```
./configure CPPFLAGS="-fPIC -I/usr/lib/openmpi/include"
    LDFLAGS="-L/usr/lib/openmpi/lib -lmpi"
```

Problem 3: If you are using **mpiP** (Vetter and McCracken, 2001) externally and during it's installation you get an error similar to this

```
libmpiP.a(wrappers.o): relocation R_X86_64_32 against `.rodata.str1.1' can not
    be used when making a shared object; recompile with -fPIC
libmpiP.a: could not read symbols: Bad value collect2: error: ld returned 1
    exit status
```

Solution: You probably need to specify the path to MPI library using this in command line when installing **mpiP**

R Script

```
./configure CPPFLAGS="-fPIC -I/usr/lib/openmpi/include"
    LDFLAGS="-L/usr/lib/openmpi/lib -lmpi"
```

Problem 4: If you are using **mpiP** (Vetter and McCracken, 2001) externally and during **pbdMPI** (Chen *et al.*, 2012a) installation you get an error similar to this

```
Error : .onLoad failed in loadNamespace() for 'pbdMPI', details:
  call: dyn.load(file, DLLpath = DLLpath, ...)
  error: unable to load shared object 'pbdMPI.so':
  pbdMPI/libs/pbdMPI.so: undefined symbol: _Ux86_64_getcontext
```

Solution: You probably need to disable some external library prerequisite by **mpiP**, using this in command line when installing **mpiP**

R Script

```
./configure --disable-libunwind CPPFLAGS="-fPIC -I/usr/lib/openmpi/include"
    LDFLAGS="-L/usr/lib/openmpi/lib -lmpi"
```

5.2 Running

Problem 1: While running **Rmpi** code for profiling, if you encounter the error below:

```
error: mpiexec was unable to launch the specified application as it could not
    access
or execute an executable:
Executable: /path/to/R/package_installation_directory/2.15/Rmpi/Rslaves.sh
Node: "Your_node"
while attempting to start process rank 0.
```

Solution: You need to make executable of the shell scripts in the "inst" directory of "Rmpi" main directory using the following command from command line in "inst" directory:

R Script

```
chmod +x *.sh
```

Problem 2: While running **Rmpi** code for profiling, if you encounter the error below:

```
[G:12221] [[39704,0],0] ORTE_ERROR_LOG: Not found in file
    ../../../../../orte/mca/plm/base/plm_base_launch_support.c at line 758
--------------------------------------------------------------------------
mpiexec was unable to start the specified application as it encountered an
    error.
More information may be available above.
--------------------------------------------------------------------------
```

Solution:

1. You need to check whether your **Rmpi** is working without the **pbdPROF**.If yes try running your **Rmpi** code on single process only.

2. If above does not help, then you may need .Rprofile in Rmpi/inst/ to run your code from "inst" directory.

3. If still your code does not run ,you need to update your OPENMPI version to the latest one.You can check your openmpi versionhttp://www.open-mpi.org/software/ompi/ through

   ```
   ompi_info
   ```

4. If further you came to this far and luck is not with you somehow(pun intended), there might some configuration problem in your machine.

6 References

Chen W-C Schmidt D, Sehrawat G, Patel P, Ostrouchov G (2013). "pbdPROF: Programming with Big Data – MPI Profiling Tools." R Package, URL http://cran.r-project.org/package=pbdPROF.

Chen WC, Ostrouchov G, Schmidt D, Patel P, Yu H (2012a). "pbdMPI: Programming with Big Data – Interface to MPI." R Package, URL http://cran.r-project.org/package=pbdMPI.

Chen WC, Ostrouchov G, Schmidt D, Patel P, Yu H (2012b). *A Quick Guide for the pbdMPI package*. R Vignette, URL http://cran.r-project.org/package=pbdMPI.

Gropp W (2000). "FPMPI-2: Fast Profiling Library for MPI." URL http://www.mcs.anl.gov/research/projects/fpmpi/WWW/.

Ostrouchov G, Chen WC, Schmidt D, Patel P (2012). "Programming with Big Data in R." URL http://r-pbd.org/.

Schmidt D, Chen WC, Ostrouchov G, Patel P (2012). "pbdBASE: Programming with Big Data – Core pbd Classes and Methods." R Package, URL http://cran.r-project.org/package=pbdBASE.

Shende SS, Malony AD (2006). "The Tau Parallel Performance System." *Int. J. High Perform. Comput. Appl.*, **20**(2), 287–311. ISSN 1094-3420. doi:10.1177/1094342006064482. URL http://dx.doi.org/10.1177/1094342006064482.

Vetter JS, McCracken MO (2001). "Statistical scalability analysis of communication operations in distributed applications." In *Proceedings of the eighth ACM SIGPLAN symposium on Principles and practices of parallel programming*, PPoPP '01, pp. 123–132. ACM, New York, NY, USA. ISBN 1-58113-346-4. doi:10.1145/379539.379590. URL http://doi.acm.org/10.1145/379539.379590.

Yu H (2002). "Rmpi: Parallel Statistical Computing in R." *R News*, **2**(2), 10–14. URL http://cran.r-project.org/doc/Rnews/Rnews_2002-2.pdf.