Parallel computing environment for R on personal cluster systems

Junji Nakano and Ei-ji Nakama

The Institute of Statistical Mathematics, Japan and COM-ONE Ltd., Japan

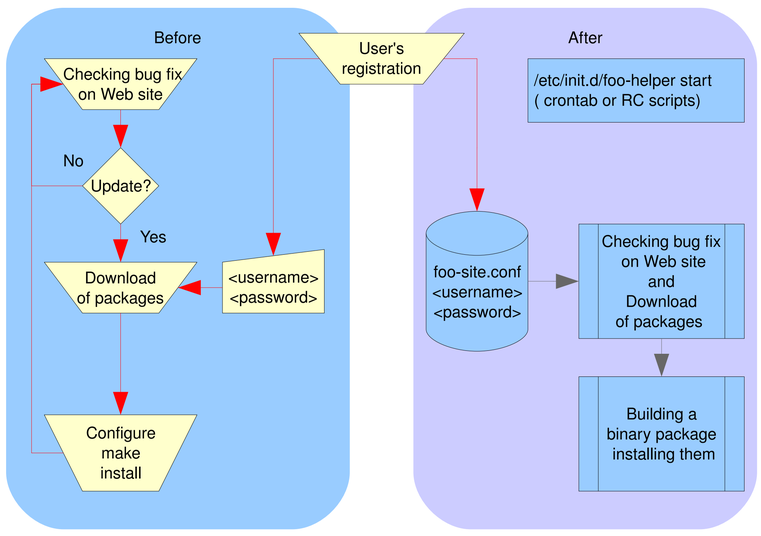
at SC09 on 14-20 November 2009 in Portland, Oregon, USA

Abstract

R is a free software environment for statistical computing and graphics. It has been used widely by both theoretical statisticians and applied statisticians working in various fields.
In order to handle large amount of data and complicated computer intensive statistical techniques, several parallel computing functions have been developed in R. These functions require parallel computing environment, which is, however, not easy to be prepared.
We have developed Debian/GNU linux helpers and packages for a job scheduler and parallel computing libraries to make their installation work easy.

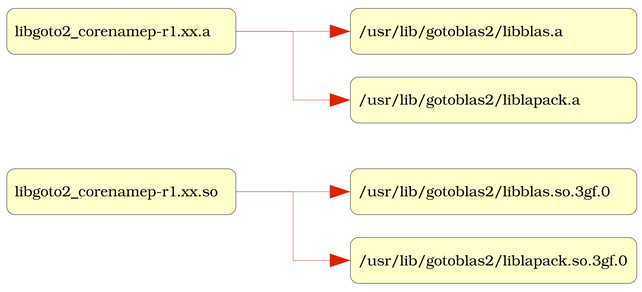
Introduction

We usually use R on supercomputers at ISM. We also have a personal cluster system (with 8 dual-core Xeons) on which Debian/GNU Linux runs. As we had some difficulties to install required libraries for parallel functions of R, we have developed helpers and packages to help their installation.
Especially, several libraries are not allowed to be distributed in binary format and need to be downloaded with user registrations. Our helpers automated user inputs in these procedures.



Helpers and packages

- torque-helper (Required)
  Torque is a resource manager for controling batch jobs and distributed computing nodes. This helper generates symbolic link to libpbs.so used in MPI library.
- maui-helper (Recommended)
  Maui is a job scheduler for clusters. It supports an array of scheduling policies, dynamic priorities, etc.
- liblam4-pbs(Required)
  LAM/MPI is a (stable) MPI implementation. This package replaces orginal liblam4 by the modified version of liblam4 to be used with Torque.
- libopenmpi1-pbs(Available as an alternative to liblam4-pbs)
  Open MPI is an ongoing project to implement MPI. This package add modules mca_pls_tm, mca_ras_tm of Torque to openMPI.
- gotoblas2-helper(Recommended)
  GotoBLAS2 is an package which includes Lapack3.1.1 and CBLAS. In Debian/GNU Linux, however, BLAS and LAPACK are separated. We devide GotoBLAS into BLAS and LAPACK following this construction.



Installation procedures

1. Modify /etc/apt/source.list.
   (a) Add CRAN mirror site of the latest R release (Lenny):
   ```
   deb http://prs.ism.ac.jp/bin/linux/debian lenny-cran
   ```
   (b) Add the site of our packages:
   ```
   deb http://prs.ism.ac.jp/~nakama/debian lenny-ism/
   ```
2. Make apt-keys available:
   (a) For CRAN mirror:
   ```
   gpg --keyserver subkeys.pgp.net --recv-key 381BA480
   gpg -a --export 381BA480 | sudo apt-key add -
   ```
   (b) For our packages:
   ```
   gpg --keyserver subkeys.pgp.net --recv-key BCEE2435
   gpg -a --export BCEE2435 | sudo apt-key add -
   ```
3. Execute update.
   ```
   sudo apt-get update
   ```
4. Execute Helper Installation.
   ```
   sudo apt-get install torque-helper gotoblas2-helper
   ```
5. Required Information by helper packages should be written in /etc/foo/foo-site.conf, where foo should be replaced by a helper name.
   ```
   # http://www.tacc.utexas.edu/?id=402
   username = unknown # uname
   password = unknown # passwd
   accept = yes # click for Licence Accept
   version = new # 1.xx or new
   ```
6. Execute Helper.
   ```
   sudo /etc/init.d/torque-helper start
   sudo /etc/init.d/gotoblas2-helper start
   ```
7. Setup Torque.
   See http://www.clusterresources.com/products/torque-resource-manager.php
8. Install LAM/MPI.
   ```
   sudo apt-get install liblam4-pbs lam4-dev
   ```
9. Install R.
   ```
   sudo apt-get install r-base r-base-dev r-recommended
   ```
10. Install snow and Rmpi.
    Note that CRAN does not have snow_0.3-4.
    ```
    sudo R --no-save <<EOF
    download.file(
      "http://www.cs.uiowa.edu/~luke/R/cluster/snow_0.3-4.tar.gz",
      "snow_0.3-4.tar.gz")
    install.packages("snow_0.3-4.tar.gz",
                     repos=NULL,
                     lib="/usr/local/lib/R/site-library")
    install.packages("Rmpi",
                     repos=c(CRAN="http://cran.r-project.org"),
                     lib="/usr/local/lib/R/site-library")
    EOF
    ```

Performance Visualization for Parallel R

snow (Simple Network Of Workstations) is a popular package for parallel computing in R. Recent snow (after snow-0.34) has functions to visualize status of parallel computing.

Listing 1: Example

```
$ qsub pvclust.sh
```

Listing 2: pvclust.sh

```
#!/bin/bash
#PBS -q q16
#PBS -N pvclust
#PBS -o pvclust.out
#PBS -e pvclust.err
export GOTO_NUM_THREADS=1
export R_DEFAULT_DEVICE=postscript
mpirun -np 1 R CMD BATCH --no-save pvclust.R pvclust.Rout
```

Listing 3: pvclust.R

```
library(pvclust)
library(MASS)
library(snow)
data(Boston)

ps.options(family="Times", width=10, height=12,
          horizontal = FALSE, onefile = FALSE, paper = "special" )
postscript<-function(file="pvclust%d.eps",bg="cornsilk",...)
    { grDevices:::postscript(file=file,,bg=bg,...) }

slave<-16
ex.pvclust<-function(nboot)
{
    ## parallel
    st1 <- snow.time({ ### start
        cl <- makeCluster(slave, type="MPI")
        boston.pv <- parPvclust(cl, Boston, nboot=nboot)
        stopCluster(cl)
    }) ### stop
    comment(st1) <- paste("Cluster Usage by pvclust(nboot=",
                          nboot, ")",sep="")

    ## non parallel
    st2 <- snow.time({ ### start
        boston.pv <- pvclust(Boston, nboot=nboot)
    }) ### stop
    comment(st2) <- paste("Non Cluster Usage by pvclust(nboot=",
                          nboot, ")",sep="")

    xlim <- c(0,max(st1$elapsed,st2$elapsed))
    ylim <- c(0,slave)

    layout(matrix(c(1,1,2),3,1))
    par(cex=1.5)
    plot(st1, xlim=xlim, ylim=ylim, title=comment(st1))
    plot(st2, xlim=xlim, ylim=c(0,2), title=comment(st2))
}

ex.pvclust(nboot=75)
ex.pvclust(nboot=1e3)
```

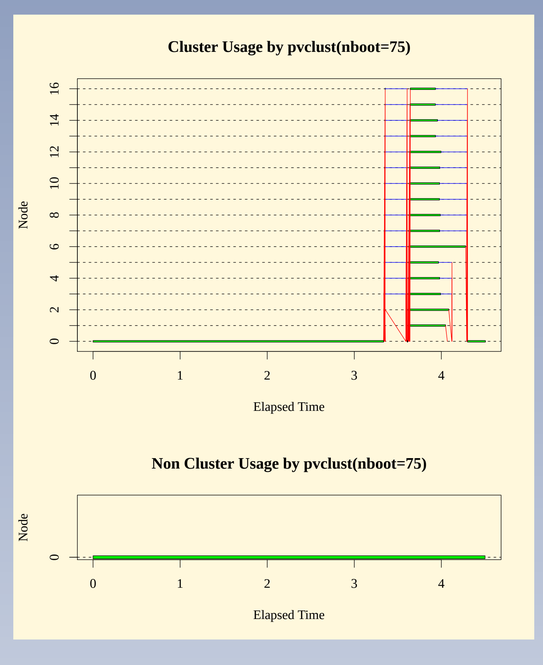


Fig 1: Small amount of parallel computing IS NOT effectve.

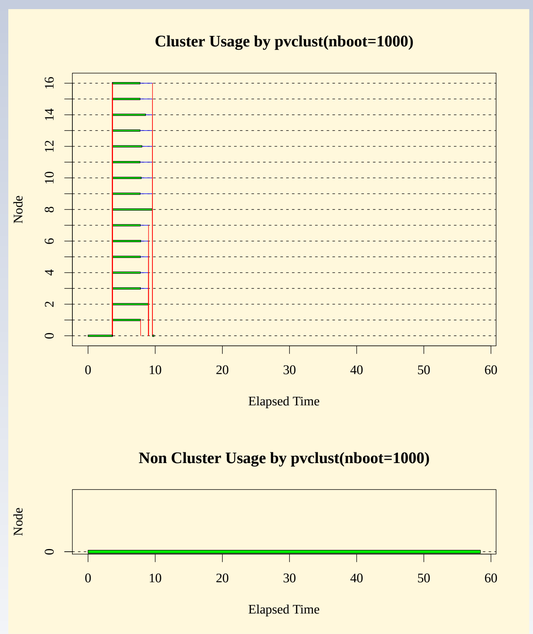


Fig 2: Large amount of parallel computing IS effective.