

# Web Interface to R for High-Performance Computing

Junji NAKANO <sup>†</sup>    Ei-ji NAKAMA <sup>‡</sup>

<sup>†</sup>The Institute of Statistical Mathematics  , Japan

<sup>‡</sup>COM-ONE Ltd., Japan

The R User Conference 2009  
July 8-10, Agrocampus-Ouest, Rennes, France



- 1 Introduction
- 2 Rdweb system
- 3 Examples of execution
- 4 Installing Rdweb
- 5 Concluding remarks



# R and requirement for huge calculation

- R: a free software environment for statistical computing and graphics for
  - statisticians to implement new statistical methods
  - practitioners to analyze real data sets in various fields
- Recently, both users require huge amount of calculation for their own purposes
- Parallel computing
  - is a practical method for realizing huge calculation
  - by executing calculations on several computers and/or many CPU cores at the same time



# Parallel computing techniques on R

- Parallel BLAS (Basic Linear Algebra Subprograms) using threads
  - ATLAS  
Free parallel and optimized BLAS
  - GotoBLAS  
Fastest parallel and optimized BLAS
  - Intel MKL, AMD ACML  
Parallel and optimized BLAS provided by vendors
- MPI type libraries for R using clustered computers
  - Rpvm  
an R interface to PVM (Parallel Virtual Machine)
  - Rmpi  
an R interface to MPI (Message Passing Interface)
- snow (Simple Network of Workstations)
  - A package for realizing parallel computing by parallel apply functions
  - Using lower level parallel libraries such as Socket, MPI, PVM, nws for transferring data among processes
  - As it conceals difference of lower level libraries, it is easy to use for parallel computing.
- multicore  
Running parallel computations in R on machines with multiple cores or CPUs.
- ...



# Existing Web environments for R

- Rweb  
A Web based interface to R for submitting the code
- Rpad  
A workbook-style user interface to R through a Web browser
- rapache  
Embedding R in the Apache Web server
- Rserve  
TCP/IP server that allows other programs to use facilities of R
- RWebServices  
Exposing R functions as Web services through Java/Axis/Apache
- ...
  - Parallel computing is not the main concern of these programs.

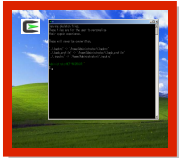


# Supercomputers in ISM

- We have three supercomputer systems in the Institute of Statistical Mathematics (ISM), Japan. (We will replace them next year.)
- Present supercomputers provide parallel computing facilities.
- We use R on our supercomputers.



# Our problems

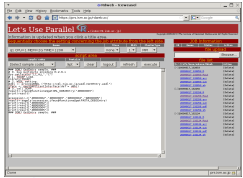


### Troubles

- Each supercomputer uses different (Unix-like) environment.
- Unix-like environments are not easy to use for novices.
- Several parameters for parallel computing need to be specified differently for each supercomputer.



# Our solution



Approach: Web interface

We have made “Rdweb”, a Web interface to R for using parallel computing functions in R

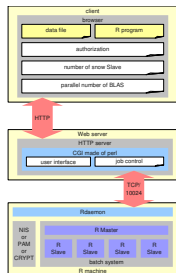
- R script edit
- file transfer
- job resource management





# Structure of Rdweb

- Rdweb (R daemon for Web) system consists of three components:
  - Web interface (via Web browser on user's computer)  
It is rather simple and programmed by HTML and JavaScript.  
JavaScript is used to assist users' input slightly.
  - Web server (on Rdweb gateway computer)  
It is a CGI program for authentication, file transfer, job control (start, stop and check), creation of JCL(Job Control Language) script and scattering the program to remote computers as a client of Rdaemon
  - Rdaemon (on the front-end computer of cluster system)  
It checks authentication, transfers required files, starts and ends jobs, and shows the status.



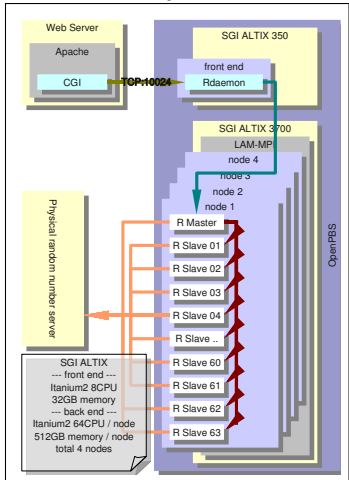
# Characteristics of Rdweb

- Rdweb is designed for supercomputers and personal PC cluster systems.
- Above stated three components of Rdweb and R slaves can reside on different or same computers.
- Text-based Web browsers can be used (with a little limitation).

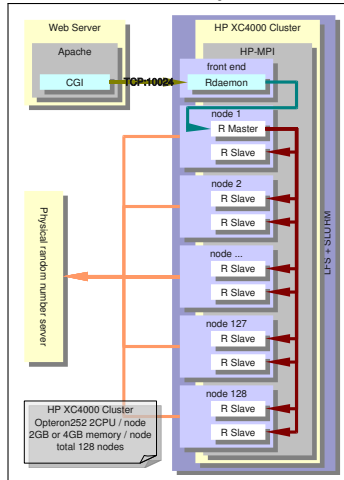


# Rdweb on supercomputers in ISM

## Shared-Memory



## Distributed-Memory



## Differences between Rweb and Rdweb

From the user side, Rdweb is similar to Rweb.

Rdweb can control system resources such as user, CPU, memory and queue. Although Rweb does not allow the use of “system” command from the security reason, Rdweb does not have such limitation because Rdweb has rigid authentication mechanism.

### Rweb and Rdweb

	Rweb	Rdweb
Authentication	none	PAM, NIS or Unix password
File upload	one file	A lot of files
Control of parallel BLAS	impossible	Each session
Control of snow	impossible	Each session

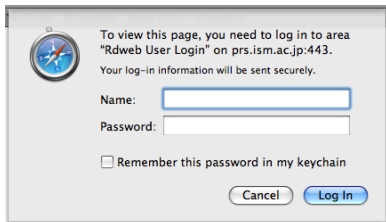


# Authentication of Rdweb (1) - Web server

Rdweb adopts two authentication stages. First stage utilizes Web server authentication mechanism when the user is connected to the Web server on the gateway computer. The mechanism is realized by `mod_auth_pam` of Apache.

## sites-enabled

```
<Directory “/www/”>
Options ....
AllowOverride None
Order allow,deny
Allow from all
AuthPAM_Enabled on
AuthType Basic
AuthName "Rdweb User Login"
Require valid-user
</Directory>
```



To view this page, you need to log in to area "Rdweb User Login" on prs.ism.ac.jp:443. Your log-in information will be sent securely.

Name:

Password:


Remember this password in my keychain

Cancel Log In



## Authentication of Rdweb (2) - Rdaemon

As second stage of Rdweb authentication, Rdaemon utilizes authentication methods such as PAM (recommended), NIS and Unix password. We can select one of them when we compile Rdweb system.



Let's Use Parallel R (ismxcr8.ism.ac.jp) Copyright:

Please input the information of the account on the super computer

Username

Password

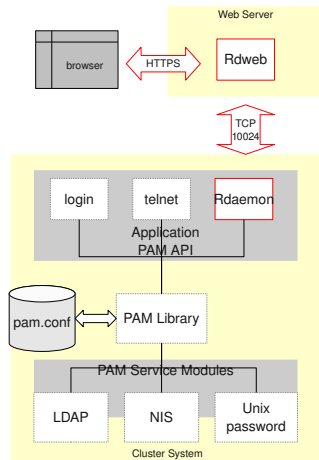
Login Cancel

Cookie must be enabled in the Web browser for Web interface of Rdweb.



# PAM authentication

- PAM (Pluggable Authentication Modules) is the API for authentication used in Linux, Solaris, MacOSX and AIX (5.3 or later).
- PAM uses NIS or LDAP or Unix password.
- If PAM is not available, NIS or Unix password can be directly used for authentication in Rdaemon.



# Location of files

“Rdweb” directory is created in the home directory on the front-end.

- Directory for execution is  
~/Rdweb/
- Uploaded files are also stored in  
~/Rdweb/
- Logs and scripts are stored in  
~/Rdweb/YYYYMMDD\_hhmmss/  
where YYYYMMDD\_hhmmss shows year, month, day, hour, minute  
and second, according to the ISO-8601 date format.





# Uploading files

- To upload data and/or program files, we click “Choose” button, select a file, and click “upload” button.
- These operations can be repeated without affecting edited script and other functions.
- SCP or SFTP clients such as Filezilla client are recommended for uploading large files because HTTP upload sometimes causes timeout and stops.

JOB Information			
ID	Queue	Status	Action
File Upload			
Choose		Upload	
File List			
[+/-] /sfs/home/nakama/Rdweb			Action



## Preparing data and program

By using a text editor, we prepare the following data file.

HW.csv

```
height,weight
1.70,65
1.85,80
1.75,86
```

Save this file as “HW.csv”.

We also prepare R program

BMI.R

```
BMI<-function(H,W)
{
    W/H^2
}
```

and save it as “BMI.R”.



# Input

Upload two files “HW.csv” and “BMI.R”. Then input the following R program

input text area

```
HW<-read.csv("HW.csv")
source("BMI.R")
HWB <- cbind(HW,BMI=BMI(HW$height,HW$weight))
HWB
plot(HWB)
```

in the editor area of Web interface which is connected to Rdweb gateway.





# Use of snow

Usually in R, we have to specify the number of processes differently according to the cluster type.

## makeCluster normal

```
# SOCK cluster  
cl <- makeCluster(c("hostname1","hostname2"))  
# MPI cluster with 2 slave processes  
cl <- makeCluster(2)
```

We add new function “setDefaultClusterOptions” to use parameters given in the Web interface in the same way for all cluster types.

## makeCluster Rdweb

```
cl <- makeCluster(getClusterOption("spec"))
```



# Selection of parameters for parallel computing

We need to select queue, number of slave processes, number of threads of parallel BLAS, and cluster type by using pull-down menus in this order.


**Sequentially choose the setting concerning the job attribute from the left side.**

Queue	Slave	BLAS	Cluster Type
q1 CPU=1 MEM=2G TIME=120H	none	1	MPI



# Execution

Job is started by clicking “Execute” button.

Let's Use Parallel  (ismxcr8.ism.ac.jp)

Information is updated when you click a title area. Copyright 2005-08 (C) The Institute of Statistical Mathematics All Rights Reserved.

**Sequentially choose the setting concerning the job attribute from the left side.**

Queue: q64 CPU=64 MEM=2G TIME=120H | slave: 31 | BLAS: 2 | Cluster Type: MPI

**Script Area**

Sample Codes: [Select Sample Code] | Font Size: 12pt | Clear | Logout | Refresh | Execute

```
require(snow)

library(MASS)
data(Boston)
library(pvclust)

system.time(boston.pv<-pvclust(Boston, nboot=100))
plot(boston.pv)

cl <- makeCluster(getClusterOption("spec"))
system.time(boston.clpv<-parPvclust(cl,Boston, nboot=100))
plot(boston.clpv)
stopCluster(cl)
```

**JOB Information**

ID	Queue	Status	Action
21079	q64	RUN	[Cancel]

**File Upload**

Choose | Upload

**File List**

Path	Action
[+/-] /sfs/home/nakama/Rdweb	
[-] 20090702_150856	[Delete]
20090702_150856.R	[Delete]
20090702_150856.Rout	[Delete]
20090702_150856.sh	[Delete]
[+] 20090702_150624	[Delete]
EVI.R	[Delete]
MW.csv	[Delete]

Creation of new result files is shown by clicking “Refresh” button.



# Batch system

Rdweb requires a batch system. Several batch systems are available.

- at, batch  
Standard batch system of Unix specified in XPG4 (X/Open portability guide Ver.4). It has simple queue mechanism.
- OpenPBS (NASA etc.)  
Queuing and scheduling control system for cluster systems. Development stopped in 1998.
- Torque (Cluster Resource Inc.)  
Free system based on Open PBS
- Load Leveler(IBM)  
Batch system by a vender
- LSF (Platform Computing Inc.)  
Commercial job controlling tool
- SLURM  
Free resource control utility





# Platforms

Rdweb should work on almost all Unix-like OSs.

We have checked the following systems in ISM and COM-ONE.

MPI	OS	BATCH SYSTEM
HP-MPI	Linux	LSF + slurm
LAM-MPI	Linux	Torque
OpenMPI	Linux	Torque
LAM-MPI	Linux	OpenPBS
LAM-MPI	Linux	at
LAM-MPI	Solaris	at
LAM-MPI	AIX	LoadLeveler
LAM-MPI	MacOSX	at

Note: Installation of these batch systems is sometimes complicated.

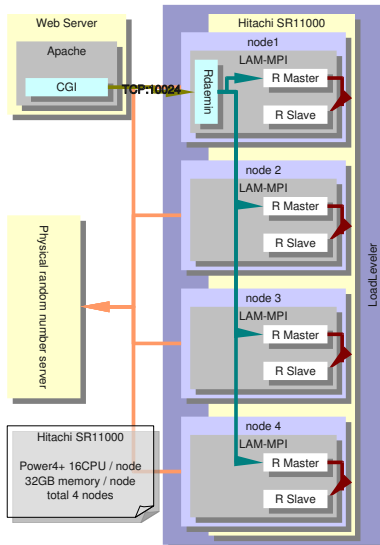
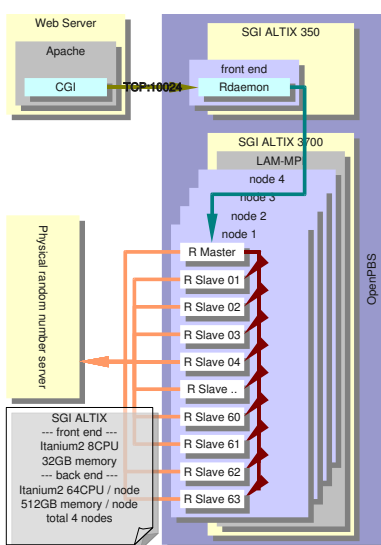


# Installation

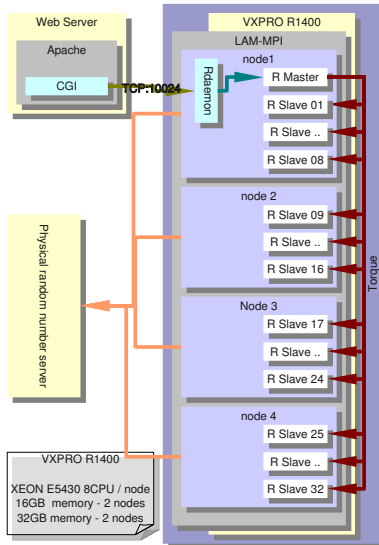
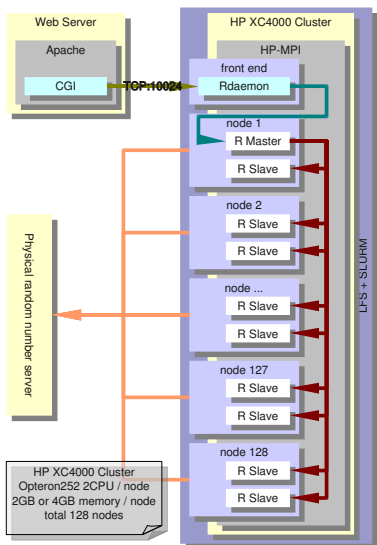
- We keep source codes of Rdweb at <http://prs.ism.ac.jp/~nakama/rdweb/>
- Required installation procedure
  - Prepare the skeleton of the shell file to a front-end
  - Define the system information on Web server
  - They depend heavily on the cluster system. Details of the setting information can be seen in “README” file in Rdweb archive.
- We put required packages for Debian GNU/Linux (Lenny) at <http://prs.ism.ac.jp/~nakama/debian/lenny-ism/>. They include helper packages for GotoBLAS, Torque, and packages of lam-mpi and openmpi for Torque. (Unfortunately, these are still buggy.)




# Examples in ISM (1)



# Examples in ISM (2)



# Concluding remarks

- Advantages of Rdweb
  - Novices can use parallel execution functions of  with less efforts.
  - Number of parallel execution can be specified easily for parallel BLAS and snow.
  - Secure authentication is available by PAM which can use LDAP or NIS.
- Disadvantages of present Rdweb
  - System installation is complicated
  - and completely platform dependent
- Future work
  - Encrypting communication between Web server and Rdaemon
  - Porting to various R
    - R with many BLASs
    - R compiled by several compilers
    - R on many OSs

