Notes on Supercomputing at OSU

Austin Carson[1]

Note: The author's experience was based on a Windows Vista platform and therefore may contain idiosyncrasies or references that are not relevant for other folks.

Note 2: Confusion results if you do not keep close attention to the two different meanings of "batch". Batch can refer to a way of using R or to the method of submitting job requests through the supercomputer's "TORQUE" program.

Website

Most information needed to get an account, use R, and use their batch processing system is available in various places at: www.osc.edu.

Setting up an account

First step is to get an account; no way to use the system without it. This is free if your research is under the name of a professor at OSU. The approval process was very quick in my experience. The goal of the application is to see why you need supercomputing resources rather than a normal computer. It requires some basic information on the project as well as a brief pitch on why the additional resources (and the amount you are requesting) are necessary. Ours was approved quickly and without questions. See:
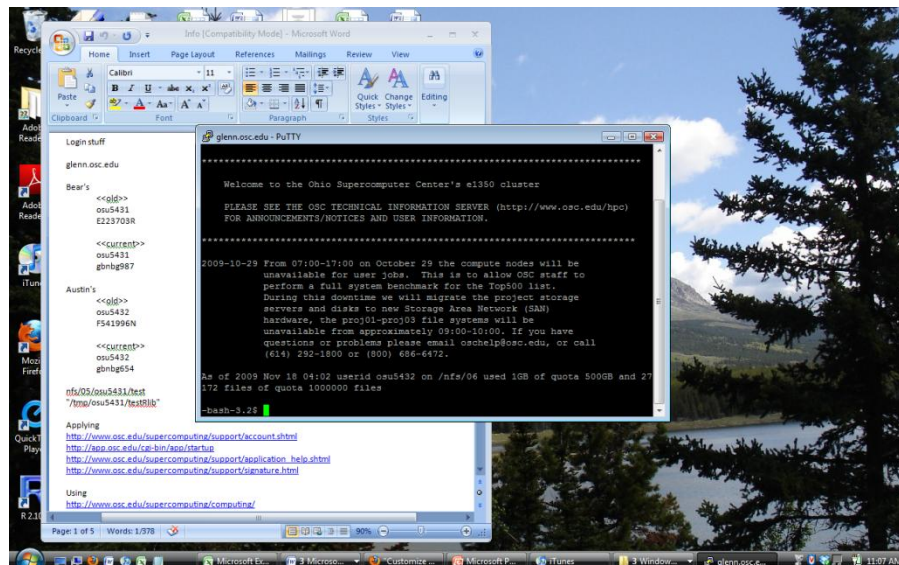
http://www.osc.edu/supercomputing/support/account.shtml
http://app.osc.edu/cgi-bin/app/startup
http://www.osc.edu/supercomputing/support/application_help.shtml

Operating environment

When logged into the supercomputer, you will be operating in Unix environment. To those unfamiliar, it looks like the following (if you are old, it looks like old DOS screens):

---

[1] These notes are based on the author's experience using the Ohio Supercomputing Center (OSC) for an OSU research project with Prof. Bear Braumoeller, July 2009-Nov 2009. Email me at carson.153@buckeyemail.osu.edu if you have specific questions.

More information on the basic commands for getting around – creating drives, copying files, seeing what files are in a folder, etc. – are at http://www.osc.edu/supercomputing/training/bunix/ and http://www.osc.edu/supercomputing/training/bunix/unix_bas_0502_pdf.pdf.

General note on use. A big part of the learning curve is getting used to syntaxes that will be foreign to many first-time users. I will try to summarize the basic components that we had to learn. Note that the supercomputer is used remotely: all users submit jobs, monitor them, and obtain the results via the use of an SSH program like PuTTy. Talk with our IT folks for help on this initial issue or read around on the internet.

Structure of the system. Once you obtain an account, you are given a "home drive" or permanent drive space which will be based on your user account number (i.e. "osu1234"). You can save files of all types there, including R script, data files, etc.[2] The basic procedure is to compose all the relevant files to execute a supercomputing job, move those files to the home drive on the supercomputer, then execute the job. It might go something like this: 1) write R batch files 2) transfer these and the relevant dataset to your home drive 3) write .job batch request 4) transfer this to your home drive. Submit a job request, let R & OSC do their thing, review results, bang head on table if there's an error, repeat.

R in batch mode. R has two ways of being used: interactive and batch. Supercomputing allows you to use either mode, but the interactive method (what we normally do) is only for very short sessions. Any lengthy processing needs to be done in R batch mode. Users have to get used to running R in this mode. This is a bit annoying because it black boxes processes we are used to being able to see and react to. It requires you to compile your list of commands in one .R or .Q document (suffixes for R script files) ahead of time, including specifying commands for the use of packages (i.e. library(foreign)) and for accessing any data files (i.e. can't "file.choose"). Then you give R one grand command (i.e. "R CMD BATCH SchultzLLS1.R") and the program crunches the numbers and spits out a single file full of results. Two tips on this front. First, it is critical that you troubleshoot and test your code before submitting it. It's very easy to have a typo or reference an object that doesn't exist or some other syntax error. Whereas normal interactive use of R spits back an error right away, batch mode will usually just seize up or give meaningless results and you may not know for awhile. Second, break apart your code. When in doubt, I took the code and broke it into four or five script files so that an error in one line did not ruin a whole job. Lastly, R output in Batch mode is saved in files that end in the suffix ".Rout" (assuming you didn't run into a fatal error that stops the process before a final output file is created). If the R script you used was "Test.R", your results will be "Test.R.Rout" and will include your code and whatever the results produced by R are. For more info on R in batch mode, see general Google search results, help files in R, and http://www.osc.edu/supercomputing/software/apps/r.shtml.

R versions. The supercomputer will have one version of R already loaded and ready to use. See http://www.osc.edu/supercomputing/software/apps/r.shtml for details. However, they do not regularly update R. We ran into problems with packages being written for a different version of R. After consulting with OSC help, we were able to install a new version of R in our home drive and use it (and those packages) without problems.

---

[2] Get a SFTP file transfer program like FileZilla to make transferring files easier (more on that below).

R packages. One early problem we ran into was the use of packages. Since the supercomputer only has a standard R version installed and has no packages installed on its system, you must manually obtain and install packages. This is made all the more difficult because of the strange operating environment (you can't just point and click using a graphical user interface because there is no GUI even when using it in interactive mode). I was able to install some packages but needed some OSC help for one of them. Note that once the packages are installed successfully, all one has to do to use them is include the library("package") command in R script files.

OSC batch job system. OSC receives requests from all kinds of folks at OSU and elsewhere. It uses a program called TORQUE to manage these job requests efficiently, distributing them to the various clusters/nodes. All requests require a strange procedure: you must create and later delete a temporary folder inside your home drive in which your job's basic elements are executed (but NOT stored).

The job request procedure requires you create a separate file (".job" created in Notepad, for me) with cryptic instructions that tell the TORQUE system basic information about your job and how to execute it. Absolutely essential reading for this is at http://www.osc.edu/supercomputing/training/customize/docs /batch/index.shtml. I have included an example of the syntax using R script called "SchultzLLS.R". The first set of instructions preceded by "#PBS" describes what type of resources your job needs: time, type of computing resources, amount of memory, etc. The second set of instructions is the meat of the request. These first tell the system to create a temporary folder ("mkdir" line), to copy the R script file to use to the temp folder, to run R in batch mode using that script file, to copy the output file into your permanent home drive, then to delete the temp folder. This takes a bit of time to get used to and is written in a Unix language which was foreign to me.[3]

```
#PBS -l walltime=80:00:00
#PBS -l nodes=1:ppn=8:oldquad
#PBS -l mem=32GB
#PBS -N Gartzketest
#PBS -S /bin/ksh
#PBS -j oe

set -x
mkdir /tmp/osu5432b
cp SchultzLLS1.R /tmp/osu5432b
cd /tmp/osu5432b

/nfs/06/osu5432/osu5432/R-2.9.2/bin/R CMD BATCH SchultzLLS1.R

cp SchultzLLS1.R.Rout $HOME
cd $HOME
rm -rf /tmp/osu5432b
exit
```

---

[3] Again, more on the basics of that are in http://www.osc.edu/supercomputing/training/bunix/ and http://www.osc.edu/supercomputing/training/bunix/unix_bas_0502_pdf.pdf.

Help

Help staff was AWESOME. Do not hesitate to email them especially. I received very useful tips on how to deal with problems (i.e. how to install some packages or use a newer version of R) and to optimize performance (i.e. how to speed up our processing). Info is:

1-800-686-6472

oschelp@osc.edu