Working with NYU HPC filesystems

So what does a job look like on NYU HPC, taking into consideration the filesystems?

- `$HOME` is a good place to keep scripts and code
- `$WORK` is a good place to keep a copy of data you will be using over the course of several months. You should have a copy of this somewhere else too, as it is not backed up
- `$SCRATCH` is for active data - this is where your jobs should run. It isn't backup up, and is periodically flushed, so make sure you have another copy of anything important.
- `$ARCHIVE` is for longer-term backup of data you are not frequently using

Below is an annotated example of how a job can use these filesystems:

```bash
#!/bin/bash
#PBS -l nodes=1:ppn=1
#PBS -l walltime=5:00:00
#PBS -l mem=2GB
#PBS -N jobname
#PBS -M bob.smith@nyu.edu
#PBS -j oe
# we run in $SCRATCH, but because it gets flushed, keep a copy of input data in $WORK
# if multiple jobs use the same input data, we might want a common "INPUT_DIR" for them:
INPUT_STORE=${WORK}/where_my_input_data_is_stored
INPUT_DIR=${SCRATCH}/where_my_job_reads_input_from
# using rsync rather than cp will copy only the files that are not already in
${INPUT_DIR}
rsync -av ${INPUTSTORE} ${INPUT_DIR}

# each job should run in a unique directory:
RUNDIR=${SCRATCH}/my_project/run-${PBS_JOBID/.*}
mkdir -p $RUNDIR
cd $RUNDIR

# load the relevant modules (in this example, amber is a chemistry/MD package)
module purge
module load amber/mvapich2/intel/14.03

# run the model
sander -O -i ${INPUT_DIR}/mdin -o my_output.dat

# if the output is important, you should copy it back to somewhere:
mkdir -p ${WORK}/results/run-${PBS_JOBID/.*}
cp my_output.dat ${WORK}/results/run-${PBS_JOBID/.*}
```

There is more about using rsync at Keeping directories in sync with rsync