Hydra-5 Migration Notes

Welcome to Hydra-5, the upgraded Hydra cluster.

1. What has changed?

a. **Hardware**
   i. 81 compute nodes for a total of 3,960 CPUs (slots, cores), some 33 old nodes have been retired.
   ii. New storage (disk space): 1.5PB GPFS high-performance parallel file system), on top of 950TB of NAS (near line storage) and 500TB of NetApp (high availability, high reliability storage).

b. **Software**
   i. Upgraded OS from CentOS 6.10 to CentOS 7.6.
   ii. Migrated scheduler from SGE 2011.11 to UGE 8.6.6 (Univa's version of the Grid Engine).
   iii. Replaced Rocks 6 with BCM 8.2 (Bright Cluster) for software management.
   iv. Change the memory reservation definition.
   v. Bioinformatics tools have been updated.
   vi. The content of the module `tools/local` has changed.

c. **Accounts**
   i. With the switch to BCM, we also switched to LDAP for user account maintenance.
   ii. We only migrated accounts of users who have logged on Hydra over the past year.

1. What has not yet changed?

a. The instructions on the Wiki (at confluence.si.edu) are a work in progress.

1. How the changes will affect users

a. **Access**
   i. You still log in via `hydra-login01.si.edu` or `hydra-login02.si.edu`, but with a new OS install, the 'RSA fingerprint' has changed, causing your next ssh to issue a warning. No need to email us, simply follow the instructions of your ssh client.

> ⚠️ Hint: you may need to delete a line in your 'known hosts' file on your own computer, which for some users is here: `~/.ssh/known_hosts`

   ii. There is no reason for you to log to the head node (`hydra-5.si.edu`), so don't; and logging on `hydra-4.si.edu` has been disabled.

b. **Accounts**
   i. With BCM we migrated to LDAP for user account management, so you will no longer need to change your password at two places. Instead, you will do it once on either login node with the passwd command.
   ii. We only migrated the ~250 users who have logged on the cluster over the past year. If you have not used Hydra for over a year, you won't have an account any longer, and you will need to contact us (Rebecca for Biology users and Sylvain for SAO users).
   iii. We heard from a couple of users who did not receive announcement emails posted to the HPCC-listserv. If you did not receive these reminders and announcements, please let Rebecca know ASAP.

c. **Passwords** ⭐
   i. Before you try to log on to the new Hydra, ***you MUST reset your password using this web page***: (https://hydra-adm01.si.edu/ssp/?action=sendtoken) Enter your hydra username in the "Login" box to receive via email a link to reset your password.
   ii. The link will be emailed to your canonical email, i.e. your @si.edu or @cfa.harvard.edu email. If you do not have one of these, it will be your institutional email (ends in .edu) if we have one on file.
   iii. This site enforces new password rules, your password will still need to be changed every 90 days. With the new self service password webpage you can reset your password if you forget your credentials or don't change your login in 90 days, so you no longer need to email us with password resets.
   iv. The other option on that web page, to use it to change your current password, is not yet working, but will be fixed soon.
   v. As mentioned above in the Accounts section, the password change process `passwd` is simplified. You only need to run this command on one of the login nodes, you no longer need to also run the command on the head node.

d. **Location of stuff and modules**
   i. We now use BCM, no longer Rocks, hence the locations of things have changed: no more `/opt`, but lots of stuff under `/cm`.
   ii. If you load modules and use environment variables, everything should work. If you hardwired a system path, you need to do the right thing (use modules and environment variables).
   iii. Because we upgraded the OS version, the versions of a lot of the standard tools have also changed too, hence things might be different or new options are now available.

e. **Submitting jobs**
   i. You still use qsub and alike, but it is now accessed via the uge module. That module is loaded for you, so unless you unload it, or do something weird, there is no need to do anything else.
   ii. The output of some GE's commands will look different, and a few take different options.
   iii. The same queues are available, except that the interactive and I/O nodes are now different nodes (our newest nodes) and the GPU and SSD queues are not yet available. These two resources have yet to be configured.

iv. Also, you will notice that the compute node naming has slightly changed: the compute nodes are all called `compute-NN-MM`, with both NN and MM are always a 2-digit string and the value of NN is shorthand for the node model (all `compute-64-MM` are all Dell R640). The fully qualified name is no longer `compute-NN-MM.local`, but `compute-NN-MM.cm.cluster`.

f. **Memory Reservation** ⭐

i. While this may be confusing and/or annoying, we have decided to change how the memory reservation is computed: ***`mres` is no longer a per slot (thread, CPU), but a per job value***.
ii. All your jobs MUST be adjusted to use the new value, which is the old value multiplied by the number of slots (threads, CPUs). h_`data` and `h_vmem` remain per slot values.
iii. You will still use the same rule as what was on hydra-4 to determine if your job should go on the high-compute or high-memory queues. If >6GB/slot is requested, the job needs to go on the himem queues: `[lmsu]THM.q`
iv. QSubGen has yet to be adjusted to reflect this change (stay tuned)!

> In other words a:
>
> ```
> -pe mthread 10 -l mres=20G,h_data=20G,h_vmem=20G,himem
> ```
>
> must be replaced by:
>
> ```
> -pe mthread 10 -l mres=200G,h_data=20G,h_vmem=20G,himem
> ```
>
> to reserve 200GB for the job.

g. **Compilers**
i. The same 3 compilers are available (gcc, icc, pgi), although the default version of the compilers has changed. You should not need to recompile your code (unless it is a Fortran code compiled with GCC).
ii. For MPI jobs, the combos of available compilers & MPI flavors has slightly changed. We hope to have this documented soon. Support for OpenMPI w/ GCC had to be dropped. Contact Sylvain if this is a problem.

h. **Bioinformatics tools** ⭐
i. Please see the wiki page for details about the Bioinformatics Software and Modules: https://confluence.si.edu/display/HPC/Bioinformatics+Software+and+Module+Information
ii. As of 9/3/2019, the bioinformatics software packages have been newly installed on Hydra-5 (with a few exceptions - those that have been transferred from Hydra-4 are noted). Note that the versions and/or module names may be different than what was on Hydra-4. An attempt was made to install the most up-to-date versions and make the module names as standardized as possible.
iii. ***Job files will need to be adjusted to reflect these changes.***
1. Users looking for software not on this list can compile/install in their own space or request installation by emailing SI-HPC@si.edu.
2. Note that with our limited time, we prioritize installing software that many users need and is well-documented.
3. Please contact SI-HPC@si.edu if you notice any issues with bioinformatics software and modules.

i. **Storage (disk space)**
i. The various disk locations are the same: `/home`, `/data`, `/pool` and `/scratch`, except that `/scratch` is now 9x bigger, faster and `/scratch/sao` is now separate from /scratch/genomics.
ii. Everyone now has a directory under `/scratch/genomics` or `/scratch/sao`.
iii. If your application is I/O intensive consider switching from using /pool to /scratch.
iv. If you already had stuff under /scratch, we have copied its content from the NetApp to the GPFS.
v. But because we had a lot to copy (120TB), we started the copying on Aug 22 at 22:54, and then rsync'd the content. This means that any file you had under `/scratch` when we started copying and have deleted afterwards might have been copied. So check your stuff if you were actively using `/scratch` after Aug 22 at 22:54.
vi. Also `/scratch` is a GPFS file system, no longer an NFS one. It should be faster, but a few commands do not work the same way, like `df` and quota.

> For the GPFS, we recommend that you use:
>
> ```
> df -h --output=source,fstype,size,used,avail,pcent,file
> ```
>
> not just
>
> ```
> df -h
> ```
>
> and to get quota information, use:
>
> ```
> module load tools/local
> quota+
> ```

vii. The command `quota+` takes options quite similar to quota (check the man page, and it will be explained on the Wiki once it is updated). Advanced users can load the mmfs module and run the GPFS specific commands
viii. We no longer allow the use of local disk space (`/state/partition1` no longer exists). Use `/scratch` instead.
ix. For those who have project-specific disk space on `/pool` and purchased `/scratch` space, we copied `/pool` to `/scratch` and the old `/pool` is mounted read-only. Contact us for more details.
x. Since we freed over 100TB on the NetApp, we will reorganize the NetApp storage: larger volumes and bigger quotas to come in the next few weeks.

j. **Local tools**
i. The tools accessible via the module tools/local have been reorganized, most have been renamed, while some have been dropped.

ii. What was loaded via tools/local was not only moved to a different location (handled by the module) but broken into tools/local and tools/local+, and the name shortened when they had an extension (like `qstat+.pl` is now `qstat+`). A few poorly chosen names have been drastically changed. Once the Wiki pages are revamped this will be explained in detail there.
iii. If you want things to be the same, load tools/local-bc (bc stands for backward compatible).

> ⚠ If you want to figure out what is what, try:
>
> ```
> module help tools/local
>
> module help tools/local+
>
> module help tools/local-bc
> ```
>
> The last one will show you what names have changed.

iv. Note that the names in some of the man pages have yet to be changed to the new names. All in due time.

k. **Cluster status page, scrubber and job monitoring**
i. The cluster status page is not yet up, but should be available within a week or so.
ii. We have suspended the scrubbing, but it will be resumed within a month or so.
iii. The job monitoring tools will be progressively restarted.