BioHPC: A High Performance Computing System for Life Sciences

Hosted Servers

Bioinformatics Facility
Biotechnology Resource Center
Cornell University

Hosted server

- Hosted server is a part of BioHPC Cloud
- BioHPC Cloud resources including storage and software are available on a hosted server the same way as on the rental servers
- Hosted server is only accessible for the hosting group associated with it
- Hosted server can be customized regarding access, software and storage.

BioHPC User Groups

- BioHPC **user** can be a member of one or more **groups**
- **BioHPC group** governs access to associated hosted server(s), their storage, other optional resources (e.g. group storage on /home)
- One member of the group should be designated as BioHPC **group manager**
- **Group manager** can add or remove group members (if they have BioHPC user IDs) and will be notified about any changes in group membership
  "My Groups" menu: https://biohpc.cornell.edu/lab/labgroups.aspx

Storage options: network storage

- Each group member has 200GB free storage in their home directory, accessible from all servers.
- A group can optionally purchase more storage shared by the group ($95 per TB-year)
- Group storage is accessible on all servers as `/home/groupname`

Storage options: local storage

- Local storage on each hosted server contains two directories: `/workdir` and `/local/storage`

- `/workdir` can only be accessed by users logged on to the server. It is never automatically cleaned on hosted servers

- `/local/storage` can be used locally or accessed from any other BioHPC server after it is mounted. It is ONLY accessible to the hosting group members.

Making local storage accessible from other servers

To mount `/local/storage` from hosted server **cbsuA** on another server **cbsuB** login to server **cbsuB** and run the the following command:

```
/programs/bin/labutils/mount_server cbsuA /storage
```

The directory will be mounted on the **cbsuB** server as `/fs/cbsuA/storage`

Controlling access to hosted servers

1. Any group member can login without reservation (most popular option),

   OR

2. Group members make reservations as needed

Multi-tasking on BioHPC servers

- Each server features multiple CPUs, GPUs (on some servers), ample memory (RAM), and local disk space

- Typically, resources are sufficient for multiple tasks (jobs), possibly from different users, to **run simultaneously**, however….

- Resources should not be oversubscribed, i.e.,

  - total number of active threads should not exceed the number of CPUs
  - combined memory consumed by all tasks should not exceed the total amount of RAM
  - there should be enough scratch disk space for all tasks

Multi-tasking on BioHPC servers

How to avoid resource over-subscription?

Option 1: direct management

be aware of CPU, RAM, disk, and timing needs of your jobs
monitor resource usage (tools like `top` may help)
communicate with other users about their jobs – not practical if group is large

Option 2: use a job scheduler (queuing system)

launch your jobs not directly, but through **scheduler's job submission tool**
specify CPU, memory, and timing of your task at submission (still need to be aware of those needs), e.g.,

```
sbatch -n 24 --mem=200000 --partition=long --time=2-13:20 script.sh
```

scheduler will

launch task when specified resources are available – otherwise keep it waiting in queue
manage job priorities between tasks and users (policies are configurable)
keep historical record (accounting) of resource usage (per user, group, etc.)

recommended for larger groups, but useful even for a single user – to balance multiple tasks

Multi-tasking on BioHPC servers

SLURM (Simple Linux Utility for Resource Management) – scheduler available for BioHPC <u>by request</u>

- (Only) Open source workload manager standard in active development
- **Efficient enforcement of CPU and memory allocations for multi-threaded processes**
- Multiple workload prioritization options
- Works with one or multiple servers
- Supports job accounting
- Does NOT control disk space or enforce disk quota
- Can control user access to servers (this feature not used; access managed by BioHPC reservation mechanisms instead)

- Can be installed on hosted servers in minutes
  - configuration has to reflect the group's needs and may take longer (days?, weeks?)

- Coming soon: possibility to use SLURM on rented machines – requested at reservation

Backup at BioHPC

https://biohpc.cornell.edu/lab/userguide.aspx?a=backupguide

Files stored on BioHPC storage (network or local) is **NOT backed up automatically**. Unless backed up, files lost due to user error (accidental deletion) or hardware crash cannot be recovered!

Optional backup can be purchased and configured via web interface https://biohpc.cornell.edu (User → My Storage).

- purchase backup storage
- configure one or more directories (such as home directory or a directory on local storage on a hosted server) to be backed up
- specify exclusions (files and subdirectories to be excluded from backup)
- specify backup frequency and the age of the oldest directory "version" to be kept on backup storage

Backup at BioHPC

https://biohpc.cornell.edu/lab/userguide.aspx?a=backupguide

How BioHPC backup works

- a current snapshot of a directory is taken each night (or at other specified frequency) around midnight, and saved on backup storage in a folder called **`current`**
- files that changed or were deleted since last snapshot have their previous copies saved on backup storage in a separate backup snapshot folder marked with the current date, e.g., **`bak_Tue_Mar_12_03:48:26_2019`**
- the **`bak_*`** snapshots older than the specified maximum age are removed from backup storage

Retrieval of data from backup

- backed up data available (read-only) on login nodes (**`cbsulogin`**, **`cbsulogin2`**, **`cbsulogin3`**) in **`/backups/backup1`** or **`/backups/backup2`** – organized by owner
- retrieval: find desired file(s) version(s) in backup folders, copy back to their original locations

Backup at BioHPC

https://biohpc.cornell.edu/lab/userguide.aspx?a=backupguide

Backup DOs and DON'Ts

DO

- plan your backup strategy carefully, consult with us if needed (brc_bioinformatics@cornell.edu)
- back up essential, hard to reproduce files (scripts, documents, raw data, results from lengthy computations)
- designate a single group member to manage group's backups

DON'T

- back up short-lived scratch files generated by your running jobs – waste of time, storage, and money!
- rename files or folders in a directory being backed up (for backup, file renamed = file deleted + another file created = backup space doubled + unnecessary backup event triggered)
- re-distribute files between subdirectories of the directory being backed up (see comment above)

Contact us before re-organizing a backup up directory!

Happy Computing !

The BioHPC Team
brc_bioinformatics@cornell.edu