Introduction to BioHPC

New User Training

[web] portal.biohpc.swmed.edu
[email] biohpc-help@utsouthwestern.edu

Updated for 2018-05-02
Overview

Today we’re going to cover:

- What is BioHPC?
- How do I access BioHPC resources?
- How can I be a good user? (some basic rules)
- How do I get effective help?

If you remember only one thing....

If you have any questions, ask us via biohpc-help@utsouthwestern.edu
What is HPC, and why do we need it?

High-performance computing (HPC) is the use of parallel processing for running advanced application programs and accessing data efficiently and reliably.

Any computing that isn’t possible on a standard PC system

**PROBLEMS**
- Huge Datasets
- Complex Algorithms
- Difficult / inefficient software

**BioHPC SOLUTIONS**
- Batch HPC jobs
- Interactive GUI sessions
- Visualization with GPUs
- Windows sessions on the cluster
- Wide range of software
- Easy web access to services
Who is BioHPC?

Liqiang Wang
   Director, 13 years experience in IT infrastructure, HPC.

Yi Du
   Computational Scientist, experience in parallel software design, large-scale data analysis.

Ross Bateman
   Technical Support Specialist, experienced in maintaining user systems and troubleshooting.

Wei Guo
   Computational Scientist, Ph.D. in Materials Science and Engineering, experience in HPC for complex simulations.

Long Lu
   Computational Scientist, MS in CS. Biology and Chemistry, gene sequencing and materials science.

Yingfei Chen
   Computational Scientist, Ph.D. in Chemical and Biological Engineering, protein structures.

Murat Atis
   Computational Scientist, Ph.D. in Physics, Computational Physics.

Li Tan
   Computational Scientist, Ph.D. in Computer Science, High Performance Computing.

Daniel Moser
   Computational Scientist, Ph.D. Atmospheric Sciences, computational fluid dynamics and cloud physics.

We are biohpc-help@utsouthwestern.edu

https://portal.biohpc.swmed.edu/content/about/staff/
What is BioHPC? - An Overview

BioHPC is:

- A 276-node heterogeneous cluster.
- >9 petabytes (9,000 terabytes) of storage across various systems.
- A large number of installed software packages.
- A network of thin-client and workstation desktop machines.
- Cloud services to access these facilities easily.
- A dedicated team to help you efficiently use these resources for your research.
**Nucleus** is our compute cluster: 11,500 CPU cores, 45 TB RAM

- **32 GB**  80 nodes
- **128 GB**  24 nodes
- **256 GB**  78 nodes
- **256 GBv1**  48 nodes with new Xeon E5v4
- **384 GB**  2 nodes

**GPU**  20 nodes/256 GB RAM/Tesla K20/K40 GPUs

**GPUv1**  12 nodes/256 GB RAM/Tesla P100 x2

**webDesktop**  4 nodes/256 GB RAM/Tesla K80 x2

Login via ssh to **nucleus.biohpc.swmed.edu** or use web portal.
**BioHPC Storage Systems**

Lysosome

- 2.5 PB DDN SFA12X
- 960 TB Dell Powervault RAID
- 35-40 GB/s
- LUSTRE Parallel Filesystem
- Connected to Nucleus with multiple Infiniti Band (56 – 100 Gbps) and 10 Gbps Ethernet links

GPFS

- 3.4 PB
- /work and /archive location
- Aggregated I/O throughput 14 GB/s (3.4 GB/s max per compute node)
As a BioHPC user, you will have access to:

- **BioHPC Cluster**
  
  - `/home2/username`  
  - `/project/department/group/username`  
  - `/archive/department/group/username`  
  - `/work/department/username`

  50 GB / user  
  5 TB / per group*  
  5 TB / per group*  
  5 TB / per user

- **BioHPC File Exchange (web-interface)**

  [https://cloud.biohpc.swmed.edu](https://cloud.biohpc.swmed.edu)  
  50 GB / user, local storage

- **BioHPC Lamella Cloud Storage (web-interface), on campus only, private cloud**

  [https://lamella.biohpc.swmed.edu](https://lamella.biohpc.swmed.edu)  
  100 GB / user, local storage

  Gateway to BioHPC Cluster (via FTP, SAMBA or WebDAV*)

* Can be increased on PI request with Dept. Chair approval.
Mirror/Full backup is the starting point for all other backups and contains all the data and files that are selected to be backed up

/home2 (Mondays & Wednesdays)
/work (Fridays)

Incremental Backup provides a faster method of backing up data than repeatedly run backups

/project (upon request)
/archive (upon request)

What data should be backed up?
How often?

http://www.backup4all.com/kb/incremental-backup-118.html
What is BioHPC? – Lamella Storage Gateway

*Lamella is our storage gateway* – two ways to access your files easily

**Web Interface / FTP Clients**

**Windows / Mac network drive mounts** *(SMB /WebDav)*

lamella.biohpc.swmed.edu
What is BioHPC? – *Thin Client & Workstation Systems*

- Desktop computers *directly connected* to the BioHPC systems.
- Login with BioHPC details, direct access to storage like on cluster.
- Run the same Linux/software packages as on the cluster, but with a graphical desktop.
- Thin client is less powerful but cheaper and smaller, compared to workstation.
What is BioHPC? - Software

A wide range of packages available on Nucleus as modules. -- two ways to check

You can ask biohpc-help@utsouthwestern.edu for additions/upgrades etc.
A big focus at BioHPC is **easy access** to our systems.

Our **cloud services** provide web-based access to resources, with only a browser.

All cloud services accessible via [portal.biohpc.swmed.edu](http://portal.biohpc.swmed.edu)
What can we do using BioHPC?

Run *any* computationally intensive work

**Linux HPC Jobs**

**GPU Visualization**

**Interactive Sessions**

**Windows with GPU Visualization**
Okay, sounds great....

But how do I use all of this?
Hands on BioHPC -- 1. Where to find all services?

portal.biohpc.swmed.edu
Quick Start - Usage Scenario

You have a complicated python script called:

hello_user.py

This script takes a few days to run on your desktop.

You join BioHPC, and now want to run this script on the cluster, so you can continue doing work on your desktop.

How would you run this script on the BioHPC cluster?

1. Upload the python script to BioHPC (Lamella)
2. Write and submit a SLURM script (Web job submitter)
Cloud storage gateway – web-based.

**https://lamella.biohpc.swmed.edu**
- 100GB separate space +
- Mount /home /project /work
- Internal

**https://cloud.biohpc.swmed.edu**
- 50GB space
- External file transfer
- Accessible from Internet
Hands on BioHPC – 2. Manage Files with Lamella / Cloud Storage Gateway

File Sharing

Lamella cloud storage [https://lamella.biohpc.swmed.edu](https://lamella.biohpc.swmed.edu) : sharing with user inside UTSW
File Exchange [https://cloud.biohpc.swmed.edu](https://cloud.biohpc.swmed.edu) : sharing with user outside UTSW
Hands on BioHPC – 2. Setting up Lamella to access project and work space

https://lamella.biohpc.swmed.edu

Log-in credentials, save in session uses the BioHPC login credentials and only saved in the user session, giving increased security. The drawbacks are that sharing is disabled, as lamella has no access to the cluster storage credentials.

Username and password mechanism requires a manually-defined username and password. Remember to click the gear icon and enable sharing.

For home leave blank
For private project space: department/lab/user
For lab shared project space: department/lab/shared
Hands on BioHPC – 2. Accessing BioHPC Storage Directly from Windows

Computer -> Map Network Drive

Folder is:  \lamella.biohpc.swmed.edu\username  (home dir)
\lamella.biohpc.swmed.edu\project
\lamella.biohpc.swmed.edu\work

Check ‘Connect using different credentials’

Enter your BioHPC username and password when prompted.
Hands on BioHPC – 2. Accessing BioHPC Storage Directly from Mac OSX

Finder -> Go -> Connect to Server

Folder is:
- smb://lamella.biohpc.swmed.edu/username (home dir)
- smb://lamella.biohpc.swmed.edu/project
- smb://lamella.biohpc.swmed.edu/work

Enter your BioHPC username and password when prompted.
Hands on BioHPC – 3. Web Job Script Generator

https://portal.biohpc.swmed.edu -> Cloud Services -> Web Job Submission
Hands on BioHPC –. 4 Web Visualization: Graphical Interactive Session via Web Portal / VNC

https://portal.biohpc.swmed.edu -> Cloud Services -> Web Visualization

Connects to GUI running on a cluster node. WebGPU sessions have access to GPU card for 3D rendering.
# Hands on BioHPC – 5. Software Modules

<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>module list</td>
<td>Show loaded modules</td>
</tr>
<tr>
<td>module avail</td>
<td>Show available modules</td>
</tr>
<tr>
<td>module load &lt;name&gt;</td>
<td>Load a module</td>
</tr>
<tr>
<td>module unload &lt;name&gt;</td>
<td>Unload a module</td>
</tr>
<tr>
<td>module help &lt;name&gt;</td>
<td>Help notes for a module</td>
</tr>
<tr>
<td>module –H</td>
<td>Help for the module command</td>
</tr>
</tbody>
</table>
Hands on BioHPC – 6. SSH Cluster Login via the Web Portal

https://portal.biohpc.swmed.edu -> Cloud Services -> Web Terminal

Connects to the login node, not a cluster node
Hands on BioHPC – 7. Connecting from Home

Windows - Follow the IR VPN instructions at:
http://www.utsouthwestern.net/intranet/administration/information-resources/network/vpn/

Mac – Try the IR instructions first. If they don’t work:

**On Campus**

Go -> Connect to Server
Server Address: smb://swnas.swmed.org/data/installs
Connect

VPN Client (Juniper) -> Juniper Mac VPN Client Installer ->JunosPulse.dmg

Install the software from in the .dmg file. You cannot test it on campus.

**At Home**

Start Junos Pulse and add a connection to server ‘utswra.swmed.edu’

When connecting must enter a secondary password, which is obtained using the ‘key’ icon in the Duo Mobile two-factor authentication smartphone app. Or type ‘push’ to get a push notification on your phone.

We can help – coffee session, or NL05.120P
**How To Be a Good User**

**HPC Systems are crowded, shared resources**

**Co-operation is necessary.**

The BioHPC team has a difficult job to do:

- Balance the requirements of a diverse group of users, running very different types of jobs.
- Make sure user actions don’t adversely affect others using the systems.
- Keep the environment secure.
- Ensure resources are being used efficiently.

Web-based Cloud-Services are designed to avoid problems.
Currently Enforced Policy

Don’t run complex things on the login node. (web terminal or nucleus.biohpc.swmed.edu)

Maximum of 16 nodes in use concurrently by any single user. 2 GPU node max per user.

Interactive use of cluster nodes using the web visualization or remoteGUI/remoteGPU scripts only*.

You cannot SSH to a computer node not allocated to you.
All we ask is…

1. If you have any question, or are unsure about something please ask us.
   biohpc-help@utsouthwestern.edu

2. When running jobs on the cluster, request the least amount of resources you
   know you need.
   Job times / memory limit / smallest node that will work etc.
   Up to a 2x margin of safety is appropriate.

3. Make reasonable attempts to use the resources efficiently.
   Run multiple small tasks on a node if you can.
   Cancel / close any jobs or sessions you no longer need.

4. Keep notes in case you need our help troubleshooting
   Keep old versions of scripts and job files
Getting Effective Help

Email the ticket system: biohpc-help@utsouthwestern.edu

What is the problem?
  Provide any error message, and diagnostic output you have

When did it happen?
  What time? Cluster or client? What job id?

How did you run it?
  What did you run, what parameters, what do they mean?

Any unusual circumstances?
  Have you compiled your own software? Do you customize startup scripts?

Can we look at your scripts and data?
  Tell us if you are happy for us to access your scripts/data to help troubleshoot.
Next Steps

• New users – register BioHPC accounts on portal and wait for activation confirmation.
• Spend some time experimenting with our systems, read the guides.
• Check the training schedule and attend relevant sessions.
• Join us for coffee sessions on last weeks of Jan., May, Aug., Oct.
BioHPC Storage – Core Users

Some core facilities provide access to BioHPC for their users to transfer data etc.

The core decides the amount and type of storage to provide to their users, e.g.

TIBIR WBMF Core:

```
/project/TIBIR/WBMF_Core/<username> 250GB* / core user
```

This is also your home directory
No separate home2 or work space

*Storage allocation and usage is at the discretion of the core, not BioHPC.*