Data Management on HCC

• Data storage on Crane
  • Different file systems and quotas

• Transferring files to and from Crane
  • Command-line file transfers (scp)
  • GUI applications for file transfers (introduction to Globus)
Data Storage: Home vs Work vs Common vs Attic

- Faster File Access (higher bandwidth / IOPS)
- Ease of Access
- Diminishing Redundancy (backups)

Attic
- Globus Access
- Additional storage for fee
- Not mounted to clusters
- Off-site Backups

Home
- Medium Speed Access
- Individual quotas (20 GB)
- Mounted on single cluster
- Daily backups

Common
- Fast Access
- Group quotas (30 TB)
- Additional storage for fee
- Mounted on all clusters
- Not backed up

Work
- Fastest Access
- No paid access
- Group quotas (50 TB)
- Mounted on single cluster
- Not backed up
- Subject to purge
Data Storage - Attic

Attic

- Data backed up in multiple locations (Lincoln and Omaha)
- Allocations for purchase
- Not mounted on the clusters
- Accessible through Globus Connect
- Intended as near-line data archive
# Data Storage – Home vs. Work vs. Common

<table>
<thead>
<tr>
<th><strong>$HOME</strong></th>
<th><strong>$WORK</strong></th>
<th><strong>$COMMON</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>- /home/[group]/[username]</td>
<td>- /work/[group]/[username]</td>
<td>- /common/[group]/[username]</td>
</tr>
<tr>
<td>- 20GB / user</td>
<td>- 50TB / group</td>
<td>- 30TB / group (larger quotas for purchase)</td>
</tr>
<tr>
<td>- Read-only on worker nodes (no active I/O for running jobs)</td>
<td>- Intended for fast I/O for running jobs</td>
<td>- Mounted on both clusters (Crane and Rhino)</td>
</tr>
<tr>
<td></td>
<td>- Short-term scratch space</td>
<td>- NOT intended for fast I/O for running jobs</td>
</tr>
<tr>
<td></td>
<td>- 6-month purge policy – not backed up!</td>
<td>- Accessible to worker nodes if you check out the license: #SBATCH -l licenses=common</td>
</tr>
<tr>
<td></td>
<td>- Examples: input/output data files for running jobs</td>
<td>- NO purge policy but not backed up!</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Examples: items that need to be accessed from both clusters such as reference databases or shared data files</td>
</tr>
</tbody>
</table>
Avoid large numbers of files

Storage resources are finite
  • Be judicious about the data you store.

Backup valuable data
Sensitive and Protected Data

HCC currently has *no storage* that is suitable for **HIPAA** or other **PID** data sets. Users are not permitted to store such data on HCC machines.
Exercises

1. Log into Crane and find out your group’s disk usage:
   $ hcc-du

2. Find out what files are scheduled to be purged (dormant for 24 weeks) from your $WORK directory: (use space bar to page through results and `q` to quit)
   $ hcc-purge -l

3. Cd to your $WORK directory, display your current path, and list the contents:
   $ cd $WORK
   $ pwd
   $ ls

4. Repeat Exercise 3 above for your $HOME and $COMMON directories.
Transferring Files – Command Line

**scp**

- Mac *terminal* and Windows 10 (version 1903 and later) *Command Prompt*

- Usage:
  
  `scp user@host:source_file user@host:target_file_or_dir`

- Example:
  
  `scp my_file.txt demo01@crane.unl.edu:/work/demo/demo01/`

- For transferring files between your computer and the cluster, always **run the scp command on your computer** (not on the cluster)
Exercises (1&2 are optional)

1. Use scp to copy a file from your computer to your work directory on Crane:

   $ scp my_file.txt demo01@crane.unl.edu:/work/demo/demo01/

2. Copy the job-examples/README.md file from your work directory on Crane to your computer:

   $ scp demo01@crane.unl.edu:/work/demo/demo01/job-examples/README.md .

3. You can use the `-r` option to scp (recursively) an entire directory. Copy the job-examples/matlab directory from your work directory on Crane to your work directory on Rhino.

   $ scp -r $WORK/job-examples/matlab demo01@rhino.unl.edu:/work/demo/demo01/
Transferring Files – GUI Applications

• **Transfer files using an SCP client**
  • WinSCP ([http://winscp.net](http://winscp.net)) – Windows
  • Cyberduck ([http://cyberduck.io](http://cyberduck.io)) – MacOS
  • Filezilla ([http://filezilla-project.org](http://filezilla-project.org)) – Windows, MacOS and Linux

• **Globus Connect ([http://globus.org](http://globus.org))**
  • Fast, secure and robust transfers with user-friendly web interface
  • Uses the High-Speed transfer nodes by default
  • Can transfer directly between clusters, Attic and personal machine
  • Other features: file sharing, rsync transfers, command-line tool (globus-cli)
  • [https://hcc.unl.edu/docs/handling_data/data_transfer/globus_connect/](https://hcc.unl.edu/docs/handling_data/data_transfer/globus_connect/)
Exercises

1. Log into globus.org using your My.UNL credentials

2. Transfer the job-examples directory from Crane to Rhino
   • Go to the File Manager tab and search for the hcc#crane and hcc#rhino endpoints.
     You will have to activate each endpoint using your HCC credentials – expires every 7 days
   • Enter the source and target paths (/work/[group]/[username])
   • Select the directory to transfer (job-examples)
   • Start the transfer

3. Download and install the Globus Connect Personal application. Transfer a file from your computer to Crane.
   • In the File Manager tab, click in the Collection text box. Then click on “Install Globus Connect Personal.”
   • Select the image for your operating system, download and follow installation instructions.
   • In the File Manager tab in the Globus web portal, search for the endpoint you just created.
   • Choose a file from your computer and transfer it to your work directory on hcc#crane.
Workflow Tips

• Test/Develop your workflow on a truncated set of data
  • Use tutorial data if provided
  • Otherwise, make your own

• Run commands in an interactive job first

• Try different combinations of SBATCH options to find what works best
  • Run multiple jobs on your truncated data to see which run faster

• Always check output/error files - even for apparently successful jobs
  • Were they created?
  • What do they contain?

• Check what resources the job actually used
  • `sacct -j <job_number> --format Elapsed,MaxRSS`
What to do if you’re stuck

• Read the HCC Documentation
  • [http://hcc.unl.edu/docs](http://hcc.unl.edu/docs)
    • If the documentation doesn’t answer your question, leave a comment or email to let us know!
  • Use `man` or `--help`
  • Look at your output and error files!!

• Consult Google
  • Useful for application specific errors or general usage

• Contact Us
  • Open Office Hours: [http://hcc.unl.edu/OOH](http://hcc.unl.edu/OOH)
  • Email [hcc-support@unl.edu](mailto:hcc-support@unl.edu)