3.0 Basic Tasks

Category: New User Orientation

Columbia Phase Out:

As of Feb. 8, 2013, the Columbia21 node has been taken offline as part of the Columbia phase out process. Columbia22-24 are still available. If your script requires a specific node, please make the appropriate changes in order to ensure the success of your job.

This section provides basic information on logging into the HPC systems, submitting and running jobs, managing jobs, storing data, transferring files. It also contains links to numerous articles that you may want to explore to get an expanded overview of the NAS environment and set-up processes.

If you have not yet obtained your default password and activated your RSA SecurID fob, please see "What You Need to Get Started" in Section 1.0, Introduction.

3.1 Logging in to the HPC Systems

Once you have obtained your default password and activated your RSA SecurID fob, you can connect to the supercomputers.

Before you log in for the first time, you may want to browse through the following articles to get an overview of the hardware and software environment and steps involved:

- Pleiades Configuration Details
- Columbia Hardware Overview
- Software Overview
- Porting & Developing: Overview

Your subsequent logins to the supercomputers can be done using one of two methods; both require use of your fob and connection through the SFEs, which are strictly for communication with systems in the enclave. Use of the SFEs should be limited to **`ssh`** or SSH Passthrough.

- Two-step connection: From your desktop to the SFEs and then to resources within secure enclave; this is quicker at first, but takes more time in the long run
- One-step connection with SSH Passthrough, the recommended method, even though it requires a bit more work to initially configure and to maintain

IMPORTANT: With either of these methods, make sure you have a Secure Shell (SSH) that supports Protocol 2 on your local desktop system. If you don't, download free OpenSSH connectivity tools.

TIP: OpenSSH 5.1 and above contains performance enhancements that could double your file transfer speeds over the WAN. You can check your version of OpenSSH by running the command **`ssh -V`** on your system. If it is older than version 5.1, an upgrade to 5.2 is highly recommended.
The examples below show how to connect to the HPC systems, using sfe1 as the secure front-end.

Conventions

- The percent sign (%) indicates the system prompt; your system prompt may be different
- **`Fixed-font`** type indicates commands, specific system names
- *Italicized* words or phrases in the example commands indicate variables and places where you substitute your own phrases, for example:

    - Substitute your NAS username for ***`nas_username`***
    - State the pfe*N* you want [*20-27*] (if you do not set up your **`.ssh/config`** file to use the Pleiades front-end load balancer)

3.1.1 Two-Step Connection Method

IMPORTANT: The secure front-end systems are used only as gateways to the enclave resources and are not to be used for storing data.

```
Local% ssh nas_username@sfe1.nas.nasa.gov
sfe1%  ssh pfe20
```

3.1.2 One-step Connection Method

To use the faster, one-step method to connect to the enclave, you must first set up SSH Passthrough and the **`.ssh/config`** file on your local system. This takes more time initially, but saves time in the long run.

IMPORTANT: If you have different usernames at the NAS facility and at your own site and want to use SSH Passthrough, you must add the following line to your **`.ssh/config`** file on your localhost if you want it to take effect for each host listed in the config file and if you do *not* want to use the command **`ssh nas_username@pfe20`**.

```
User nas_username
```

or

```
host pfe20 pfe20.nas.nasa.gov
  ProxyCommand ssh nas_username@sfe1.nas.nasa.gov /usr/local/bin/ssh-proxy %h
```

For your convenience, we provide a downloadable .ssh/config template called **config_nas.txt**. After you download this file, copy the file to your **.ssh** directory, replace <NAS_login_name> with your actual NAS login name in the file, and then rename the file **config**.

See the security overview for links to additional related topics.

3.2 Submitting and Running Jobs

Now that you've done all the basic setup (and assuming your allocation is in place) you're ready to run jobs.

Computer allocations are specified in Standard Billing Units (SBUs). Note that when your allocation is expended, you will no longer be able to run jobs, and you'll need to request more hours. You can check your remaining SBU balance by issuing the command **acct_ytd** on the system(s) where you have accounts.

```
% acct_ytd
```

3.2.1 Usage Charge Methods

Below is a simple explanation of how you are charged for system usage.

- **Front-End Systems**
  Usage of the HPC front-end nodes (the bridge nodes, PFEs, and cfe2) is *not* charged. However, remember that these front-end nodes are meant for editing and/or compiling and running short testing jobs. If you misuse these systems, your jobs will be terminated.
- **Back-End Systems**
  Usage of the back-end compute nodes *is* charged. The number of Standard Billing Units (SBUs) charged to a job is the number of total wall-clock hours used multiplied by the Minimum Allocation Units (MAUs) and the following SBU rates:

Pleiades (Sandy Bridge)	1.82
Pleiades (Westmere)	1
Pleiades (Nehalem-EP)	0.8
Pleiades (Harpertown)	0.45
Columbia	0.18

3.2.2 Understanding Queues

All NAS facility supercomputers use the Portable Batch System (PBS) to manage both interactive and batch jobs. The available queues on different systems vary, but all typically have constraints on maximum wall-time and/or the number of CPUs allowed for a job. Some queues may also have other constraints or be restricted to serving certain users or project ID numbers (GIDs). In addition, mission directorate limits are also set on the supercomputers. For more information, see the Pleiades Mission Shares Policy.

The table below lists commands to view various types of queue information.

% qstat -Q **% qstat -Qf**	available queues and their constraints on a system
% qstat -W shares=-	mission shares for each mission on a system

3.2.3 Interactive Processing

All jobs running on the back-end nodes, either through an interactive session or a batch session, are done by issuing the **qsub** command to PBS, from the PFEs or bridge nodes for Pleiades, as described below.

To run a job interactively, use the command **qsub -I** (as in Interactive) in conjunction with the resources you request via the **-l** (as in "liquid") option. For example, if you were on pfe20, to request 1 node with 8 CPUs and 1 wall-clock hour on Pleiades:

```
pfe20% qsub -I -lselect=1:ncpus=8,walltime=1:00:00
```

TIP: When you issue commands for Columbia, you will likely use the **qsub** command on cfe2, although c21, c22, c23, c24 will also work.

For Columbia, a similar resource request, for example from cfe2, is:

```
cfe2% qsub -I -lncpus=8,walltime=1:00:00
```

3.2.4 Batch Processing

Before submitting a batch job, you must first create a PBS job script. Below is a sample script where the **select** option is used to request 4 nodes, 8 cores per node (indicated by **ncpus**) and one hour of maximum elapsed time. Without an explicit request of a queue name, the job uses the default run queue.

```
#PBS -l select=4:ncpus=8
#PBS -l walltime=1:00:00
```

```
cd $PBS_O_WORKDIR

mpiexec -np 32 ./a.out
#end of script
```

To submit this job to Pleiades, for example from bridge3, issue the following command:

```
bridge3% qsub name_of_your_sample_script
```

3.2.5 Charging to a Non-default GID

For users who have access to more than one GID, only one of those GIDs is set as your default, which is listed in the **/etc/passwd** file(s) of the system(s) you have access to. Your jobs will be charged to the default GID unless you specify a different one.

If you want to charge your usage to a non-default GID (for example, s0901), add the following to your PBS script:

```
#PBS -W group_list=s0901
```

3.2.6 Checkpointing

None of the HPC operating systems have automatic checkpoint capability. For jobs that need a lot of resources and/or long wall-time, you should implement a checkpoint/restart capability in the source codes or job scripts.

PBS automatically reruns unfinished jobs after system crashes or certain system issues. If you do not want PBS to rerun your job, make sure to add the following in your PBS script:

```
#PBS -r n
```

3.3 Managing Jobs

When a job is submitted to PBS, a job identification number (JOBID) is assigned. PBS jobs are usually in the Running (R) state or the Queued (Q) state. Use the following commands to find status and manage your jobs; using JOBID 12345 as an example:

%qstat	show the status of batch jobs
%qstat -nu *your_nas_username*	list all jobs that belong to you
%qstat 12345 or	check the status of a PBS job using the JOBID

%qstat -f 12345	
%qstat -s 12345	show why a job is not yet running
%qdel 12345	delete a job
%qhold 12345	place a job on hold

Again, for jobs that need lots of resources and/or long wall-time, you should implement a checkpoint/restart capability in the source codes or job scripts.

See also: Common Reasons for Being Unable to Submit Jobs and Common Reasons Why Jobs Won't Start.

3.4 Storing Data

For each HPC system you have access to, you are given the following:

- A home directory to store a small number of files such as source code, small input files, etc.
- A /nobackup (scratch) directory to use for reading and writing large amounts of data while running jobs

For long-term data storage, you also have access to a home filesystem on one of the Lou mass storage systems.

3.4.1 Filesystem Overview

To find out if your home filesystem is on Lou1 or Lou2, log in to either of them, and run the command **mylou**. In the example below, your Lou home directory is on Lou2:

```
Lou2% mylou
Your Mass Storage host is lou2
```

Store files there in your home directory, **/u/*your_nas_username***

For Pleiades Lustre filesystems (called /nobackup), the actual disk that your /nobackup directory resides on is one of /nobackupp[*1-6*]. A sym-link is created so that you can simply type **/nobackup/*your_nas_username*** to access your /nobackup directory.

Also note that the Pleiades /nobackup filesystems are mounted on Lou2. So, you can get access to data on these filesystems from the Lou system without logging in to any Pleiades hosts. The also can be accessed from the bridge nodes (bridge[1-4]).

For Columbia, there are two types of /nobackup directories. One is a local /nobackup for

each Columbia host that can *only* be accessed from a backend host (such as /nobackup[*21-24*] for Columbia[*21-24*]). The other is a CXFS filesystem (that is, /nobackup[*1-2*][*a-i*]).

To find out which Pleiades /nobackup you are assigned to on Pleiades, type:

```
% ls -ld /nobackup /your_nas_username
```

3.4.2 File Retention Policy (backup and nobackup)

All home filesystems (including the Lou systems) are backed up each night. These backups are stored for approximately one year.

WARNING: The **`/nobackup`** file systems mean just that: they are *not* backed up. While this is stating the obvious, some users have lost important data by storing it on these systems. It is your responsibility to copy valuable data to either your home directory, to archival storage on the Lou systems, or to your remote system.

3.4.3 Mass Storage

The HECC environment includes two mass storage systems Lou1 and Lou2 (115 PB total storage capacity). These systems allow you to retrieve files quickly and securely. Data stored on disk is migrated to tapes as needed to make space for more data. Two copies of your data are written to two separate tape media in silos located in two different buildings.

NAS currently has no specified quota for mass storage space, but a limit of 250,000 files can be stored.

Data Migration Facility (DMF)

The Lou systems automatically write and retrieve data to and from tape. The data migration (disk-to-tape) and recall (tape-to-disk) are managed by SGI's Data Migration Facility (DMF).

You can log into the Lou systems like any other system and save data to mass storage by simply copying your files to your home directory on any of the Lou systems, referenced as **`lou`*`X`*`:/u/`*`your_userid`*** where *X* is 1 or 2. However, you will only have storage space on one of them. Use the command **`mylou`** to find out which Lou system you should store your data on.

You can manually copy files from your home filesystems (or any of the /nobackup filesystems of Pleiades or Columbia) to Lou for long-term storage (disk and/or tape archive).

See Data Migration Facility Commands for more information.

Storage Commands

Here are a few tips to make your mass storage transactions more efficient.

- Instead of using the command **ls** to list your files, use the command **dmls** to see whether the files are on tape (OFL) or on disk (REG or DUL)
- Use the command **du** with the **--apparent-size** option to see how much data is in a directory, as shown below (substituting your directory name for *datadir*):

```
lou2 % du -sh --apparent-size datadir
150G    datadir
```

- Use the command **dmget** to retrieve a set of files from tape all at once, as shown below:

```
lou2 % dmget data.200905* &
```

3.5 Transferring Files

This section provides information and available file transfer commands you can use to transfer files either between your desktop and the HPC systems or to transfer files between the NAS HPC systems. See File Transfer Overview for more information.

3.5.1 File Transfer Commands

The table below summarizes file transfer commands, any pros and cons, and performance numbers.

Command	Use to	Benefits & Drawbacks	Performance
cp	Copy between any mounted filesystems of a host		
mcp	Copy between two mounted filesystems of a host	Over **cp**: multi-threaded can be significantly faster than **cp** on both CXFS and Lustre filesystems	Up to 1.8 gigabytes per second for large files (2+ GB)
cxfscp	Copy between two mounted filesystems of a host; at least one of the filesystems	Over **cp**: Uses multiple threads and large direct I/Os to get full bandwidth (about 4-7 times faster than **cp** for files over 2 gigabytes (GB))	Up to 400 megabytes per second (MB/sec)

	is CXFS		
	Transfer files between Lou systems and Pleiades		
scp	Secure copy between two hosts; both authentication information and data are encrypted Typically for transferring small files within the NAS facility (< 5 GB) or offsite (<1 GB)		Within the secure enclave: 40-100 MB/sec, depending on various factors For faster **scp**, use HPN-SSH or upgrade to OpenSSH 5.1 or newer
bbftp	File transfer between two hosts; authentication information is encrypted, data is unencrypted Transferring large files (> 1 GB) within NAS or offsite	Over **scp**: Can transmit multiple streams of data; possibly better transfer rate Complicated syntax	Within the secure enclave: similar or better than **scp** (40-100 MB/sec) Over WAN: up to 50MB/sec (may reach 100 MB/sec to certain NASA sites)
bbscp	When you want bbFTP functionality but with SCP-like syntax; in-house wrapper script for bbFTP	Installed on all HPC systems (under **/usr/local/bin**)	Same as **bbftp**
shiftc	Transfers files between two hosts or between file systems on the same host	Transfer tracking and notification, automated failure recovery, automatic transport selection/optimization, and optional integrity verification. External remote transfers must	Up to 1.8 GB/s for local copies on single host. Up to 400 MB/s between Pleiades bridge nodes and Lou. Up to 100 MB/s between other secure enclave systems. Up

	be initiated from external host via SUP. Recommended.	to 100 MB/s over high speed WAN links. Even faster performance possible when using automatically parallelized transfers. Note that maximum performance depends on various factors.

We encourage you to explore our new file transfer tool, Shift, for performing reliable local, enclave-to-enclave, and remote transfers.

3.5.2 Transfering Files Between Enclave and Outside

For inbound file transfers that are initiated from your local system to HPC systems, you will be prompted for (at minimum, depending on your SSH Passthrough setup) your SecurID when you issue the transfer command.

Outbound file transfers that are initiated from the HPC systems to your local hosts may be a better way to transfer files, if your local hosts allow it.

Below are four examples of initiating **scp**, **bbftp**, **bbscp**, and **shiftc** from your localhost, assuming that you have different usernames between your local system and the NAS systems. Substitute your own filenames for *foo* or *filename*.

Example 1:

```
Local% scp nas_username@lou2.nas.nasa.gov:foo ./foo
```

Example 2:

```
Local% bbftp -u nas_username -e 'setnbstream 2; get filename' -E
'bbftpd -s -m 2' lou2.nas.nasa.gov
```

Example 3:

```
Local% bbscp foo nas_username@lou2.nas.nasa.gov:
```

Example 4:

```
Local% sup -u nas_username shiftc lou2.nas.nasa.gov:foo./foo
```

3.5.3 Transfer Files Within the Enclave

To transfer files between supercomputers, you can use the commands **cp**, **mcp**, **cxfscp**, **scp**, **bbftp**, **bbscp**, or **shiftc** (see table of commands, above). Most users transfer files among three systems: Pleiades, Columbia, and Lou.

Transfer Files Between Pleiades and Columbia Filesystems

Since the Columbia CXFS filesystems are mounted on the Pleiades bridge nodes, you can simply log in to one of the bridge nodes, bridge[*1-4*], and use the **`cp`** or **`shiftc`** commands to do the file transfer.

Transfer Files Between Pleiades and Lou

Pleiades /nobackup are mounted on Lou2, enabling disk-to-disk copying, which should give the highest transfer rates. You can use **`shiftc`**, **`cp`**, or **`mcp`** to copy files or make tar files directly from Pleiades /nobackup to your Lou home directory.

Network file transfers to the Lou systems will go over the 10 GigE interface by default. The commands **`shiftc`**, **`scp`**, **`bbftp`**, and **`bbscp`** are available to do file transfers; **`bbscp`** uses mostly the same syntax as **`scp`**, but performs faster. **`shiftc`** is recommended as it will use the highest performing transports and will automatically handle the various aspects of transfers to/from the Lou tape-backed file systems.

File transfers from the compute nodes to Lou must go through the front-end nodes (PFEs and bridge nodes) first.

TIP: When sending data to Lou systems, please keep your largest individual file size under 1 TB, as large files will keep all of the tape drives busy, preventing other file restores and backups.
If you use the commands **`bbftp`** or **`bbscp`** to transfer files to Lou systems, limit the number of streams to 2. Using a higher number of streams may overwhelm the Lou filesystems when multiple **`bbscp`** or **`bbftp`** connections are in process.

Transfer Files between Columbia and Lou

To transfer files between Columbia and Lou, use the commands **`shiftc`**, **`scp`**, **`bbftp`**, or **`bbscp`**.

Article ID: 264
Last updated: 14 Feb, 2013
New User Orientation -> 3.0 Basic Tasks
http://www.nas.nasa.gov/hecc/support/kb/entry/264/?ajax=1